

Pensare senza parole

Il substrato subpropositivo della cognizione e i suoi limiti formali

Proemio: Il privilegio indebito del linguaggio

La storia della filosofia della mente e della scienza cognitiva è attraversata da un'assunzione che raramente viene esplicitata proprio perché è così pervasiva da sembrare ovvia: che il pensiero sia essenzialmente linguistico, o che almeno le sue forme più alte e articolate lo siano. Questa assunzione ha radici profonde — in Aristotele che definisce l'uomo come *ζῷον λογικόν*, nel Tractatus di Wittgenstein in cui i limiti del linguaggio sono i limiti del mondo, nella svolta linguistica del Novecento che ha fatto del linguaggio il medium privilegiato di ogni analisi filosofica. Ha anche una giustificazione pragmatica irresistibile: il pensiero che possiamo studiare, comunicare e analizzare è quello che si è già tradotto in forma linguistica. Il resto — il pensiero prima e sotto le parole — rimane opaco agli strumenti ordinari dell'analisi.

La tesi di questa monografia è che questo privilegio del linguaggio sia indebito e che le sue conseguenze siano gravi, sia per la teoria della cognizione sia per la comprensione di ciò che i modelli di linguaggio fanno e non fanno. Il pensiero subpropositivo — il pensiero che avviene nelle immagini mentali, nei pattern motori, nell'intuizione esperta, nelle strutture tonali e ritmiche, nei modelli spaziali — non è il residuo grezzo che precede il pensiero vero, in attesa di essere tradotto in proposizioni. È un sistema computazionale primario con proprietà formali proprie, con una struttura di rappresentazione distinta dal linguaggio naturale, con capacità che il pensiero linguistico non ha e — crucialmente — con limiti che il pensiero linguistico supera, ma a un costo che la teoria standard non contabilizza.

L'argomento si sviluppa in cinque movimenti. Il primo costruisce il modello formale del substrato subpropositivo come sistema di rappresentazione distribuita in uno spazio di attivazione ad alta dimensione — un sistema strutturalmente diverso dal linguaggio proposizionale e non riducibile ad esso. Il secondo analizza le proprietà computazionali specifiche di questo sistema: cosa può fare che il pensiero linguistico non può, e viceversa. Il terzo affronta il processo di traduzione dal subpropositivo al proposizionale — la codifica linguistica del pensiero — come compressione lossy con struttura specifica, e le distorsioni sistematiche che introduce. Il quarto estende l'analisi ai modelli di linguaggio, mostrando che operano esclusivamente nel dominio proposizionale e che questo li priva strutturalmente del substrato da cui la cognizione umana più efficace attinge. Il quinto affronta il limite: i domini in cui il substrato subpropositivo è irriducibile e il pensiero linguistico è fondamentalmente inadeguato.

Capitolo I — Il substrato subpropositivo come sistema di rappresentazione distribuita

I.1 La geometria degli spazi di attivazione

Il punto di partenza è neurobiologico, ma la sua formalizzazione è matematica. I neuroni corticali operano in reti di milioni di unità, e ogni stato cognitivo — ogni percezione, ogni ricordo, ogni pensiero — corrisponde a un pattern di attivazione distribuito su quelle reti. Lo stato cognitivo non è localizzato in nessun singolo neurone o insieme piccolo di neuroni: è un vettore in uno spazio ad alta dimensione, dove ogni coordinata corrisponde al livello di attivazione di una singola unità.

Si definisce formalmente lo spazio di attivazione cognitivo come $\mathbb{R}^n$, dove n è il numero di unità nel sistema. Ogni stato cognitivo è un vettore $v \in \mathbb{R}^n$. La metrica naturale su questo spazio è la distanza euclidea, ma per la cognizione è più rilevante la distanza angolare — il coseno dell'angolo tra i vettori:

$$\text{sim}(v_1, v_2) = (v_1 \cdot v_2) / (\|v_1\| \cdot \|v_2\|)$$

che misura la similarità semantica tra stati cognitivi indipendentemente dalla loro intensità. Due stati cognitivi con $\text{sim} \approx 1$ sono strutturalmente simili — attivano le stesse unità nelle stesse proporzioni. Due stati con $\text{sim} \approx 0$ sono strutturalmente ortogonali — attivano unità diverse.

I.2 La rappresentazione distribuita come substrato del pensiero subpropositivo

Il pensiero subpropositivo opera direttamente in $\mathbb{R}^n$ — produce trasformazioni di vettori di attivazione senza passare attraverso una rappresentazione simbolica discreta. Le operazioni fondamentali del substrato subpropositivo sono:

L'interpolazione: dati due stati v_1 e v_2 , il sistema può produrre stati intermedi $\alpha \cdot v_1 + (1-\alpha) \cdot v_2$ per $\alpha \in [0,1]$. Questa operazione non ha equivalente nel pensiero proposizionale: le proposizioni non si interpolano — non esiste una proposizione che sia "a metà strada" tra "è rosso" e "è blu". Ma le rappresentazioni distribuite si interpolano naturalmente, producendo stati che corrispondono a sfumature, transizioni, gradi — la struttura dell'esperienza continua che il linguaggio discretizza.

La composizione per sovrapposizione: dati due stati v_1 e v_2 , il sistema può produrre $v_1 + v_2$ — la sovrapposizione che attiva simultaneamente le strutture di entrambi. La sovrapposizione non è la congiunzione logica " v_1 e v_2 " — che è una proposizione discreta — ma la coattivazione di due strutture in un unico stato. Il pensiero metaforico — "Giulio Cesare è un leone" — non è la congiunzione di due

proposizioni ma la proiezione dello spazio di attivazione di "leone" nello spazio di attivazione di "Giulio Cesare": la sovrapposizione che produce un terzo stato con proprietà di entrambi.

L'analogia come trasformazione lineare: se $v_A : v_B :: v_C : v_D$, allora esiste una trasformazione lineare T tale che $T(v_A) \approx v_B$ e $T(v_C) \approx v_D$. L'analogia — il meccanismo cognitivo centrale nella scoperta scientifica e nella creatività — è una trasformazione dello spazio di attivazione, non un'inferenza proposizionale. Il celebre esempio dell'embedding word2vec — "re - uomo + donna $\approx$ regina" — non è un'analogia proposizionale ma una trasformazione lineare in $\mathbb{R}^n$.

I.3 La struttura del pensiero proposizionale come proiezione discreta

Il pensiero proposizionale — il pensiero in linguaggio naturale o in qualsiasi altro sistema simbolico discreto — è una *proiezione* dello spazio continuo $\mathbb{R}^n$ su un insieme discreto di categorie. Ogni proposizione corrisponde a una regione di $\mathbb{R}^n$ — l'insieme degli stati di attivazione compatibili con quella proposizione — e la codifica linguistica è la mappa che associa a ogni vettore v la proposizione la cui regione lo contiene.

Formalmente, sia L l'insieme delle proposizioni del linguaggio. La funzione di codifica linguistica è:

$$\Phi: \mathbb{R}^n \rightarrow L, \quad \Phi(v) = \operatorname{argmin}_{\{p \in L\}} d(v, R_p)$$

dove $R_p \subset \mathbb{R}^n$ è la regione dello spazio di attivazione associata alla proposizione p e $d(v, R_p)$ è la distanza del vettore v dalla regione. La codifica è la proiezione al rappresentante proposizionale più vicino.

Questa proiezione è necessariamente multi-a-uno: molti vettori v diversi vengono mappati nella stessa proposizione p . La perdita di informazione nella proiezione è:

$$I_{\text{lost}} = H(V) - H(V \mid \Phi(V))$$

l'entropia della variabile aleatoria sullo spazio di attivazione meno l'entropia condizionale dato il codice linguistico. Questa quantità misura quanto del pensiero subpropositivo non è catturato dalla sua codifica linguistica: la complessità della struttura del pensiero che il linguaggio appiattisce.

I.4 La misura della perdita subpropositiva

Si può stimare I_{lost} per diverse categorie di pensiero. Per il pensiero proposizionale standard — il ragionamento su categorie ben definite, la manipolazione di relazioni logiche — I_{lost} è bassa: la struttura del pensiero è già essenzialmente discreta, e la codifica linguistica è quasi senza perdita. Per il pensiero percettivo — il riconoscimento di volti, la valutazione estetica di una composizione musicale, la

navigazione spaziale — I_{lost} è alta: la struttura del pensiero è continua e multidimensionale, e la codifica linguistica perde gran parte della finezza strutturale.

La situazione estrema è quella dell'intuizione esperta — il pensiero del giocatore di scacchi che "vede" la mossa giusta, del medico che riconosce la diagnosi prima di elaborarla, del matematico che ha "la sensazione" che un approccio sia corretto. Questi stati cognitivi hanno struttura ricca in $\mathbb{R}^n$ — sono vettori altamente informativi che discriminano finemente tra possibilità — ma si traducono malissimo in linguaggio. L'esperto che cerca di spiegare la propria intuizione produce spesso una razionalizzazione post hoc — una proposizione che è nel vicinato della regione R_p corrispondente al vettore v dell'intuizione, ma che non cattura la struttura fine che rende l'intuizione corretta.

I.5 La dimensionalità intrinseca del substrato subpropositivo

Un risultato importante della teoria delle varietà manifold è che i dati ad alta dimensione spesso giacciono su varietà a dimensionalità intrinseca molto più bassa. Uno spazio di attivazione con $n = 10^6$ neuroni non ha effettivamente 10^6 dimensioni di variazione rilevante: la struttura cognitiva rilevante giace su una varietà intrinsecamente di dimensione $d \ll n$.

La *dimensionalità intrinseca* del substrato subpropositivo per un dominio cognitivo specifico è:

$$d_{int} = \dim(M)$$

dove $M \subset \mathbb{R}^n$ è la varietà che gli stati cognitivi rilevanti abitano. Per il riconoscimento visivo dei volti, studi empirici sulla geometria degli spazi di attivazione corticale suggeriscono $d_{int} \approx 50-100$, molto inferiore agli $n = 10^8$ neuroni della corteccia visiva. Per il linguaggio, d_{int} è dell'ordine di 10^2-10^3 — la struttura semantica del linguaggio è bassa-dimensionale rispetto alla ricchezza potenziale dello spazio neurale.

La dimensionalità intrinseca ha conseguenze dirette sulla compressione. La codifica linguistica con un vocabolario di $|L|$ proposizioni può catturare al massimo $\log_2 |L|$ bit di struttura. La struttura del substrato subpropositivo ha entropia $H(V) \approx d_{int} \cdot \log_2(1/\epsilon)$ bit, dove ϵ è la precisione di discretizzazione. Il rapporto:

$$Comp = (d_{int} \cdot \log_2(1/\epsilon)) / \log_2 |L|$$

misura il fattore di compressione della codifica linguistica: quanto più ricco è il substrato rispetto alla sua espressione linguistica. Per la percezione visiva, $Comp$ può essere dell'ordine di 100-1000: il linguaggio cattura l'1% della struttura del pensiero percettivo.

I.6 Il pensiero motorio come sistema computazionale primario

Tra i sistemi subpropositivi, il pensiero motorio merita attenzione separata per la sua struttura computazionale propria. Il sistema motorio non è soltanto un esecutore di istruzioni linguistiche: è un sistema di rappresentazione e pianificazione che opera in uno spazio di traiettorie, forze, posture e coordinazioni che non ha rappresentazione linguistica adeguata.

Si formalizza il pensiero motorio come operante nello spazio delle traiettorie nello spazio del compito. Sia $\tau: [0, T] \rightarrow \mathcal{Q}$ una traiettoria nello spazio della configurazione $\mathcal{Q}$ del sistema motorio. Il "pensiero" del sistema motorio è una distribuzione su traiettorie:

$$p(\tau \mid \text{intenzione, contesto})$$

che seleziona le traiettorie compatibili con l'intenzione motoria nel contesto corrente. Questa distribuzione è ad alta dimensione — $\mathcal{Q}$ ha tipicamente dimensione 100-1000 per il sistema corpo umano — e la sua struttura è radicalmente diversa da qualsiasi struttura proposizionale.

Il pensiero motorio dell'esperto — il pianista, il chirurgo, l'atleta — ha la proprietà che la distribuzione $p(\tau \mid \text{intenzione, contesto})$ è concentrata su traiettorie di alta qualità: l'esperto "sa" come muoversi prima di elaborare linguisticamente la propria conoscenza. Questa conoscenza procedurale è irreducibile al linguaggio non perché sia ineffabile in senso mistico ma perché la sua struttura è intrinsecamente continua e ad alta dimensione: qualsiasi codifica linguistica perde la finezza che rende la traiettoria esperta superiore a quella del novizio.

I.7 La relazione tra substrato subpropositivo e linguaggio: non subordinazione ma interfaccia

Il punto fondamentale che il quadro formale stabilisce è che il rapporto tra substrato subpropositivo e linguaggio non è di subordinazione — il substrato non è il pensiero grezzo in attesa di essere elevato alla forma linguistica — ma di *interfaccia*: due sistemi con strutture computazionali distinte che si interfacciano attraverso le funzioni di codifica Φ e decodifica Φ^{-1} .

La funzione di codifica $\Phi: \mathbb{R}^n \rightarrow L$ traduce il substrato in linguaggio. La funzione di decodifica $\Phi^{-1}: L \rightarrow \mathbb{R}^n$ traduce il linguaggio in substrato. Nessuna delle due è invertibile: Φ è multi-a-uno (perde informazione), e Φ^{-1} è ambigua (la stessa proposizione può corrispondere a stati subpropositivi diversi). L'interfaccia introduce perdite in entrambe le direzioni: la codifica linguistica perde struttura subpropositiva, e la decodifica linguistica produce stati subpropositivi che non sono univocamente determinati dal testo.

Questa struttura di interfaccia ha conseguenze precise per l'analisi dei modelli di linguaggio. Un modello che opera esclusivamente su L — che non ha accesso a $\mathbb{R}^n$ — non è soltanto privo del substrato subpropositivo: è privo dell'intera struttura che la funzione di decodifica Φ^{-1} produce quando un umano legge il testo. Quando un

umano legge "il profumo della pioggia sulla terra asciutta", Φ^{-1} produce un vettore $v \in \mathbb{R}^n$ che attiva strutture olfattive, tattili, mnestiche — la ricchezza sensoriale che il testo evoca. Il modello di linguaggio non produce questo vettore: produce una distribuzione sulle parole successive compatibili con quel testo, calcolata interamente su L . La perdita non è di intensità ma di tipo: manca intera dimensione computazionale.

Capitolo II — Proprietà computazionali del substrato subpropositivo

II.1 Il repertorio computazionale del substrato

Il capitolo precedente ha stabilito che il substrato subpropositivo è un sistema di rappresentazione distribuita in $\mathbb{R}^n$ strutturalmente distinto dal linguaggio proposizionale. Questo capitolo analizza il repertorio computazionale specifico di ciascun sistema — ciò che ciascuno può fare meglio dell'altro — con l'obiettivo di produrre una mappa delle capacità cognitive che è più informativa della semplice dicotomia "conscio/inconscio" o "verbale/non verbale".

La mappa ha la struttura di una matrice di competenze: per ogni tipo di operazione cognitiva, si indica quale sistema — subpropositivo, proposizionale, o la loro interfaccia — è computazionalmente più efficiente.

II.2 Operazioni in cui il substrato eccelle: il riconoscimento di pattern

Il riconoscimento di pattern è l'operazione in cui il substrato subpropositivo ha il vantaggio computazionale più netto. Dato un vettore v di input, il substrato deve determinare a quale categoria cognitivamente rilevante appartiene — quale sia il pattern più simile tra quelli precedentemente appresi.

Nel substrato, questo calcolo è:

$$\text{cat}(v) = \underset{k}{\operatorname{argmax}} \{ \text{sim}(v, \mu_k) \} = \underset{k}{\operatorname{argmax}} \{ (v \cdot \mu_k) / (\|v\| \cdot \|\mu_k\|) \}$$

dove μ_k sono i prototipi delle categorie apprese. Il calcolo è $O(K \cdot n)$ dove K è il numero di categorie e n è la dimensione dello spazio: lineare nel numero di categorie, realizzabile in parallelo attraverso le connessioni sinaptiche. Il riconoscimento di un volto tra miliardi di volti possibili avviene in meno di 200 millisecondi non perché il cervello sia "veloce" in senso astratto, ma perché il calcolo della somiglianza nello spazio di attivazione è realizzato in parallelo da miliardi di sinapsi simultaneamente.

Il pensiero proposizionale non può fare altrettanto. Il riconoscimento di pattern nel dominio proposizionale richiede il confronto esplicito tra la descrizione dell'input e le descrizioni di tutte le categorie possibili — un processo seriale con complessità $O(K \cdot |\text{descrizione}|)$. Per K grande e descrizioni ricche, questo è computazionalmente intrattabile nel limite delle risorse cognitive reali. Il sistema subpropositivo risolve il problema di riconoscimento in un modo fondamentalmente diverso: non attraverso il confronto proposizionale ma attraverso la similarità geometrica nello spazio di attivazione.

II.3 La generalizzazione geometrica: estrapolare dalla struttura dello spazio

Una conseguenza strutturale della rappresentazione distribuita è la generalizzazione geometrica: la capacità di rispondere correttamente a istanze mai viste, sulla base della posizione che occupano nello spazio di attivazione rispetto alle istanze note.

Si formalizza la generalizzazione geometrica come segue. Il sistema ha appreso un insieme di coppie input-output $\{(v_i, y_i)\}$. Dato un nuovo input v , il sistema produce l'output:

$$y(v) = \sum_i w_i(v) \cdot y_i, \text{ dove } w_i(v) \propto \text{sim}(v, v_i)$$

una media pesata degli output noti, con pesi proporzionali alla similarità con l'input nuovo. Questo è il meccanismo dei metodi kernel — il kernel di Nadaraya-Watson — e corrisponde all'interpolazione geometrica nello spazio di attivazione.

La generalizzazione geometrica ha proprietà che la generalizzazione proposizionale non ha. In particolare, produce output continui per input continui: piccole variazioni dell'input producono piccole variazioni dell'output. Questo è il meccanismo dell'intuizione esperta: il maestro di scacchi vede una posizione leggermente variata rispetto a posizioni note e produce una valutazione interpolata — non un'inferenza proposizionale su regole ma un'interpolazione geometrica nella propria esperienza precedente.

II.4 La composizionalità del substrato: non riducibile alla composizionalità proposizionale

Il pensiero proposizionale ha una proprietà che viene spesso citata come suo vantaggio strutturale fondamentale: la *composizionalità sistematica*. Il significato di una proposizione è funzione composizionale del significato delle sue parti: se si conosce il significato di "cane" e di "corre", si può comprendere "il cane corre" senza averlo incontrato prima, perché le regole di composizione sono sistematiche e produttive.

Il substrato subpropositivo ha una forma diversa di composizionalità che non è riducibile alla composizionalità proposizionale: la *composizionalità per sovrapposizione*. Dato il vettore $v_{\text{cane}} \in \mathbb{R}^n$ che rappresenta "cane" e il vettore $v_{\text{corre}} \in \mathbb{R}^n$ che

rappresenta "corre", il substrato produce un vettore composito che è una combinazione non lineare di v_{cane} e v_{corre} — una sovrapposizione che attiva le strutture di entrambi in modo mutuamente modulato.

La differenza è strutturalmente importante. La composizionalità proposizionale è *discreta e produttiva*: può generare espressioni arbitrariamente complesse da un vocabolario e una grammatica finiti. La composizionalità per sovrapposizione è *continua e contestuale*: produce stati che dipendono dalla geometria specifica dei vettori componenti e dal contesto dell'attivazione. I due meccanismi non sono intercambiabili: la composizionalità proposizionale è superiore per la generazione di espressioni strutturalmente nuove, mentre la composizionalità per sovrapposizione è superiore per la produzione di stati semanticamente sfumati, contestualmente modulati, analogicamente ricchi.

Formalmente, si può dimostrare che:

$$KC(\Phi^{-1}(p \wedge q)) \leq KC(v_p + v_q)$$

la complessità del vettore prodotto dalla composizionalità proposizionale è inferiore a quella della sovrapposizione subpropositiva — il pensiero proposizionale paga meno in termini di KC per la composizione — ma:

$$\dim(M_{\{v_p + v_q\}}) > \dim(M_{\{\Phi^{-1}(p \wedge q)\}})$$

la dimensionalità intrinseca della varietà di stati compatibili con la sovrapposizione supera quella degli stati compatibili con la congiunzione proposizionale — la sovrapposizione produce stati più ricchi strutturalmente.

II.5 La temporalità del substrato: sequenza, ritmo e struttura dinamica

Un aspetto del substrato subpropositivo che il pensiero proposizionale può descrivere ma non replicare computazionalmente è la struttura temporale fine. Il substrato opera in tempo reale — le sue trasformazioni sono processi dinamici che si svolgono nella dimensione temporale — e la struttura temporale è parte integrante del contenuto cognitivo.

Si formalizza la struttura temporale del substrato come traiettoria nello spazio di attivazione:

$$v(t): [0, T] \rightarrow \mathbb{R}^n$$

La cognizione musicale è il caso paradigmatico: la comprensione di una melodia non è la comprensione di un insieme di note — è la comprensione della traiettoria $v(t)$ nello spazio di attivazione tonale. La struttura della traiettoria — le accelerazioni, le tensioni,

le risoluzioni, i pattern ritmici — costituisce il contenuto cognitivo musicale. Questa struttura non è proposizionale: nessuna descrizione proposizionale di una melodia la sostituisce come oggetto cognitivo.

Il pensiero proposizionale può descrivere la struttura temporale — "la melodia sale di una quinta e poi risolve per semitono" — ma questa descrizione è la codifica $\Phi(v(t))$, non la traiettoria stessa. La perdita nella codifica è massima per le strutture temporali con alta variazione fine: le nuances di timing che distinguono un'esecuzione musicale esperta da una meccanica, i micro-movimenti che costituiscono la competenza atletica, il ritmo prosodico che porta l'emozione nel parlato.

II.6 L'esperienza qualitativa come struttura del substrato

Il problema difficile della coscienza — perché l'esperienza soggettiva ha qualità fenomeniche — è naturalmente irrisolto da qualsiasi teoria puramente funzionale. Ma il quadro del substrato subpropositivo permette di precisare dove il problema si colloca rispetto alla struttura computazionale.

I *qualia* — le qualità dell'esperienza soggettiva, il rosso del rosso, il dolore del dolore — sono tipicamente descritti come irriducibili a qualsiasi descrizione funzionale. Nel quadro formale, questa irriducibilità ha una struttura precisa: i qualia sono proprietà dei vettori $v \in \mathbb{R}^n$ che non sono preservate dalla proiezione proposizionale Φ .

Formalmente, due vettori v_1 e v_2 possono soddisfare $\Phi(v_1) = \Phi(v_2)$ — avere la stessa descrizione proposizionale — pur essendo strutturalmente distinti: $\text{sim}(v_1, v_2) < 1$. Questa distanza non proposizionale tra stati con la stessa descrizione linguistica è la struttura formale dell'indistinguibilità proposizionale degli zombie filosofici: due sistemi con la stessa descrizione funzionale proposizionale possono occupare posizioni diverse in $\mathbb{R}^n$, e quella differenza posizionale — irrilevante per Φ — è dove i qualia risiedono se risiedono da qualche parte.

Questo non risolve l'hard problem: non spiega perché la posizione in $\mathbb{R}^n$ produca esperienza soggettiva. Ma lo localizza con precisione: il problema non è nella struttura proposizionale della cognizione, dove tutti i funzionalisti cercano, ma nella struttura subpropositiva che la proiezione proposizionale scarta.

II.7 Operazioni in cui il pensiero proposizionale eccelle: la produttività ricorsiva

Il quadro non deve produrre una rivalutazione acritica del substrato a scapito del linguaggio. Il pensiero proposizionale ha vantaggi computazionali genuini che il substrato non può replicare.

Il più importante è la *produttività ricorsiva*: la capacità di generare strutture arbitrariamente complesse attraverso l'applicazione ricorsiva di regole di composizione. Il substrato subpropositivo può produrre sovrapposizioni di un numero limitato di componenti — la sua capacità composizionale è limitata dalla

dimensione dello spazio $\mathbb{R}^n$ e dalla linearità del meccanismo di sovrapposizione. Il pensiero proposizionale può generare strutture ricorsivamente profonde — proposizioni su proposizioni su proposizioni — senza limite di principio sulla profondità.

Formalmente, la *profondità ricorsiva* del pensiero proposizionale è illimitata:

$$\text{depth}(p) = \max \{ d : p \text{ contiene una sottostruttura di profondità } d \}$$

può essere arbitrariamente grande. La profondità ricorsiva del substrato è bounded dalla dimensionalità intrinseca d_{int} dello spazio: non può rappresentare strutture con profondità maggiore di d_{int} senza perdita di struttura.

Questa asimmetria spiega perché il linguaggio è stato evolutivamente vantaggioso per l'uomo: il pensiero proposizionale ricorsivo permette di pianificare sequenze di azioni arbitrariamente lunghe, di ragionare su stati ipotetici annidati — "se lui crede che io penso che lui voglia..." — di costruire teorie con struttura logica profonda. Questi compiti richiedono profondità ricorsiva che il substrato non può fornire da solo.

II.8 Il confine ottimale tra substrato e linguaggio: una funzione di costo

Si può ora definire formalmente il confine ottimale tra pensiero subpropositivo e pensiero proposizionale per un dato compito cognitivo. Il confine è determinato dal rapporto tra il costo computazionale di ciascun sistema per quel compito e la qualità della soluzione che produce.

Sia T un compito cognitivo. Il costo del substrato è:

$$\text{Cost}_{\text{sub}}(T) = \text{KC}(\text{soluzione}_T) / d_{\text{int}} \quad (\text{per unità di dimensione})$$

il costo per dimensione della varietà del substrato rilevante per T . Il costo del linguaggio è:

$$\text{Cost}_{\text{lang}}(T) = \text{KC}_{\{\text{lang}\}}(\text{soluzione}_T) \quad (\text{complessità proposizionale})$$

la complessità di Kolmogorov della soluzione nel linguaggio proposizionale. Il sistema ottimale usa il substrato per le parti del compito in cui $\text{Cost}_{\text{sub}} < \text{Cost}_{\text{lang}}$ e il linguaggio per le parti in cui $\text{Cost}_{\text{lang}} < \text{Cost}_{\text{sub}}$.

Il confine ottimale produce una *divisione del lavoro* cognitivo: il substrato gestisce il riconoscimento di pattern, l'interpolazione geometrica, la struttura temporale fine; il linguaggio gestisce la pianificazione ricorsiva, il ragionamento esplicito, la comunicazione con altri agenti. I sistemi cognitivi più efficaci — gli esperti in qualunque dominio — sono quelli che hanno imparato a usare ciascun sistema per le

parti del compito in cui eccelle, e a interfacciarli fluentemente attraverso le funzioni Φ e Φ^{-1} .

Capitolo III — La codifica linguistica come compressione lossy del substrato

III.1 Il problema della traduzione interna

Ogni volta che un agente cognitivo produce linguaggio — parla, scrive, formula un pensiero in parole — esegue una traduzione interna: trasforma un vettore $v \in \mathbb{R}^n$ in una sequenza di simboli $p \in L$. Questa traduzione è la funzione di codifica Φ introdotta nel Capitolo I, e come ogni codifica lossy introduce distorsioni sistematiche. Il problema della traduzione interna non è studiato sistematicamente dalla linguistica o dalla psicologia cognitiva — che tendono a trattare il linguaggio come lo strumento trasparente attraverso cui il pensiero si esprime — ma è il problema centrale per comprendere perché il pensiero difficile è difficile da comunicare, perché la comprensione profonda precede e supera la sua formulazione verbale, e perché certi domini sono resistenti alla formalizzazione proposizionale.

III.2 La struttura delle distorsioni della codifica linguistica

La codifica $\Phi: \mathbb{R}^n \rightarrow L$ introduce distorsioni che non sono casuali ma hanno struttura determinata dalla topologia di L . Si identificano cinque tipi principali.

La *discretizzazione categoriale* è la distorsione più pervasiva: il continuum delle gradazioni nel substrato viene mappato su categorie discrete del linguaggio. Il continuum cromatico dello spazio di attivazione visiva viene mappato su un numero finito di termini di colore che variano da lingua a lingua — il risultato classico di Berlin e Kay sul relativismo cromatico. La struttura dell'esperienza percettiva non ha confini netti tra i colori; la codifica linguistica impone confini dove il substrato ha transizioni graduate. L'informazione persa è esattamente quella che si trova al confine delle categorie: le gradazioni che il linguaggio tratta come equivalenti ma che il substrato distingue.

La *linearizzazione sequenziale* è la distorsione prodotta dal fatto che il linguaggio è intrinsecamente sequenziale — una stringa di simboli — mentre il substrato è intrinsecamente parallelo e multidimensionale. Un vettore $v \in \mathbb{R}^n$ ha struttura simultanea: tutte le sue coordinate sono presenti contemporaneamente e si influenzano reciprocamente. La sua codifica linguistica deve serializzare questa struttura in una sequenza, scegliendo un ordine che è arbitrario rispetto alla struttura del vettore. La distorsione di linearizzazione è la perdita di informazione sulla struttura multidimensionale che la serializzazione introduce.

La *granularità lessicale* è la distorsione prodotta dal fatto che il vocabolario di L ha una granularità fissa — i termini del linguaggio hanno significati che sono più o meno grossolani rispetto alla struttura del substrato nel dominio corrispondente. In domini per cui la lingua ha vocabolario ricco — le relazioni sociali nelle lingue con molti termini di parentela, la neve nell'Inuit, le emozioni nelle lingue con alta granularità emotiva — la distorsione di granularità è bassa. In domini per cui il vocabolario è povero — le sensazioni somatiche fini, le qualità degli stati emotivi complessi, le strutture spaziali tridimensionali — la distorsione è alta.

La *struttura presupposizionale* è la distorsione prodotta dal fatto che ogni proposizione porta presupposizioni implicite — strutture del substrato che devono essere condivise tra mittente e destinatario perché la proposizione sia interpretata correttamente. Queste presupposizioni non sono codificate esplicitamente nel segnale ma devono essere ricostruite dal destinatario attraverso Φ^{-1} . Quando le presupposizioni del mittente non coincidono con le aspettative del destinatario, la decodifica produce un vettore diverso da quello del mittente — il malinteso del Capitolo I della monografia precedente come caso specifico di distorsione presupposizionale.

La *coersione aspectuale* è la distorsione prodotta dalla struttura grammaticale del tempo e dell'aspetto, che impone una prospettiva temporale sul substrato che può non corrispondere alla struttura temporale dell'evento cognitivo. Il pensiero di un evento può essere strutturato come un processo continuo nel substrato ma deve essere codificato in una delle strutture aspectuali disponibili nella lingua — puntuale, durativo, iterativo — che può non catturare la struttura del processo come vissuto.

III.3 La funzione di tasso-distorsione per la codifica linguistica

Si può ora costruire la funzione di tasso-distorsione per la codifica linguistica del pensiero subpropositivo. La distorsione è misurata dalla distanza angolare tra il vettore originale v e il vettore prodotto dalla decodifica $\Phi^{-1}(\Phi(v))$:

$$d_{\text{sub}}(v, \Phi^{-1}(\Phi(v))) = 1 - \text{sim}(v, \Phi^{-1}(\Phi(v)))$$

La funzione di tasso-distorsione è:

$$R_{\text{sub}}(D) = \min_{\Phi} \{ H(V) - H(V | \Phi(V)) \mid \mathbb{E}[d_{\text{sub}}] \leq D \}$$

il minimo dell'informazione mutua tra il substrato e il codice linguistico compatibile con distorsione media D . Per $D = 0$ — distorsione zero — si richiede che Φ sia iniettiva, il che è impossibile se $|\mathbb{R}^n| \gg |L|$ — se lo spazio di attivazione ha più stati del vocabolario linguistico. Il lower bound sulla distorsione del linguaggio come codice del substrato è:

$$D_{\text{min}} = 1 - 2^{-\{H(V) - \log_2 |L|\}}$$

per una sorgente con entropia $H(V)$ codificata con un vocabolario di dimensione $|L|$. Questa è la distorsione irriducibile del linguaggio come codice del substrato: anche il codice linguistico più efficiente non può ridurre la distorsione al di sotto di $D_{\min}$.

III.4 L'esperienza dell'inesprimibile: fenomenologia della distorsione massima

Il punto in cui la distorsione della codifica linguistica è massima corrisponde a una fenomenologia precisa che ogni umano conosce: la sensazione di avere qualcosa da dire che non si riesce a dire, di capire qualcosa che non si riesce a spiegare, di sapere come fare qualcosa senza poter insegnarlo. Questi non sono stati di confusione — sono stati di alta chiarezza subpropositiva con bassa traducibilità linguistica.

Il caso del musicista è paradigmatico. Un pianista esperto sa come suonare una certa frase con l'espressione giusta — il suo substrato motorio e musicale ha un vettore v con struttura ricca che guida l'esecuzione. Ma se gli viene chiesto di spiegare verbalmente come produce quell'espressione, la codifica $\Phi(v)$ produce una descrizione molto impoverita: "ci vuole un po' più di peso sul secondo tempo", "devi sentire la tensione armonica". Queste descrizioni sono vere ma parziali — catturano dimensioni di v ma non la struttura complessiva che rende l'esecuzione giusta.

La distorsione massima si verifica quando il dominio cognitivo ha dimensionalità intrinseca d_{int} alta e vocabolario linguistico povero — quando $R_{\text{sub}}(D_{\min}) \gg 0$ e $D_{\min}$ è grande. I tre domini di massima distorsione sono: l'esperienza percettiva fine (sapori, odori, texture), la competenza procedurale esperta (il "tocco" in qualsiasi arte o artigianato), e l'esperienza emotiva complessa (i modi sfumati del benessere e del malessere che il vocabolario emotivo standard non distingue).

III.5 Il pensiero ai margini del linguaggio: le forme transitorie di codifica

Tra il substrato subpropositivo e il linguaggio proposizionale esiste un continuum di forme di codifica con livelli intermedi di perdita. Si identificano tre livelli di codifica sub-linguistica che meritano analisi separata.

Il *pensiero in immagini* è una codifica di $v \in \mathbb{R}^n$ in uno spazio visivo strutturato — un'immagine mentale che preserva la struttura spaziale e le relazioni percettive del substrato ma le traduce in un formato analogico anziché proposizionale. La distorsione dell'immagine rispetto al substrato è inferiore a quella del linguaggio: l'immagine preserva la continuità spaziale, le gradazioni di colore, le relazioni geometriche che le proposizioni discretizzano. La distorsione dell'immagine rispetto al substrato è superiore a quella del substrato stesso: l'immagine perde la struttura multisensoriale — l'odore, il sapore, la consistenza — che il vettore v portava.

Il *pensiero in schemi motori* è una codifica di $v \in \mathbb{R}^n$ in traiettorie nello spazio del movimento — un'esecuzione o una simulazione motoria che preserva la struttura temporale e cinestesica del substrato. Il pianista che "sente con le mani" la frase

musicale sta operando con una codifica in schemi motori: il suo sistema di rappresentazione motoria preserva strutture del substrato musicale che la codifica linguistica perderebbe.

Il *pensiero in strutture tonali e ritmiche* è una codifica che preserva la struttura temporale fine del substrato attraverso pattern di attivazione ritmicamente strutturati. Il ritmo del pensiero — la velocità con cui le idee si succedono, la tensione tra strutture che si risolve — è una proprietà del substrato che influenza la qualità del pensiero finale anche dopo che quest'ultimo è stato codificato in linguaggio. I teorici della letteratura che analizzano il ritmo della prosa di un autore stanno analizzando la traccia che la struttura temporale del substrato lascia nella codifica linguistica.

III.6 La perdita progressiva nella trascrizione: dal substrato al testo scritto

Il processo di produzione linguistica ha struttura a cascata: substrato → articolazione → testo parlato → testo scritto. Ogni passo introduce distorsioni aggiuntive.

Il passaggio dal substrato all'articolazione introduce la linearizzazione e la discretizzazione categoriale analizzate nel paragrafo III.2. Il passaggio dall'articolazione al testo parlato aggiunge la struttura prosodica — intonazione, ritmo, enfasi — che porta parte della struttura temporale del substrato attraverso la modulazione acustica. Il passaggio dal testo parlato al testo scritto elimina la prosodia: la trascrizione ortografica perde l'intonazione, il ritmo, le pause, le variazioni di intensità che nella voce parlata portavano dimensioni del substrato.

Si può stimare la perdita progressiva. Sia I_{sub} l'informazione totale del substrato. La codifica in articolazione preserva $I_{\text{art}} \leq I_{\text{sub}}$. La codifica in testo parlato preserva $I_{\text{speech}} \leq I_{\text{art}}$ — la prosodia aggiunge I_{prosody} rispetto al contenuto proposizionale puro. La trascrizione in testo scritto preserva $I_{\text{text}} \leq I_{\text{speech}} - I_{\text{prosody}}$. La perdita totale è:

$$I_{\text{lost}} = I_{\text{sub}} - I_{\text{text}} = I_{\text{sub}} - I_{\text{art}} + I_{\text{art}} - I_{\text{speech}} + I_{\text{prosody}}$$

Per domini ad alta struttura subpropositiva — la poesia parlata, la comunicazione emotiva intensa, la dimostrazione pratica — I_{lost} è grande: gran parte del contenuto del substrato non è nel testo scritto finale. Per domini a bassa struttura subpropositiva — la matematica formale, la logica proposizionale — I_{lost} è piccola: il substrato e il testo coincidono quasi completamente.

Questa analisi ha una previsione verificabile rispetto alle tradizioni culturali che preferiscono la trasmissione orale alla scrittura: quelle tradizioni hanno accesso a strutture cognitive — le sfumature prosodiche, i pattern ritmici, i gesti che accompagnano il discorso — che la scrittura perde, e la loro resistenza alla trascrizione non è irrazionalità ma preservazione dell'informazione subpropositiva che il testo eliminerebbe.

Capitolo IV — I modelli di linguaggio come sistemi senza substrato

IV.1 L'architettura dell'assenza

Il quadro formale dei capitoli precedenti permette di formulare con precisione la tesi centrale di questo capitolo: i modelli di linguaggio di grandi dimensioni non sono sistemi con un substrato subpropositivo impoverito rispetto a quello umano — sono sistemi *senza substrato subpropositivo nel senso definito*, che operano esclusivamente nello spazio simbolico L e non hanno accesso a $\mathbb{R}^n$.

Questa affermazione richiede una qualificazione importante. I modelli Transformer hanno internamente rappresentazioni distribuite ad alta dimensione — gli embedding dei token sono vettori in $\mathbb{R}^d$ con d tipicamente dell'ordine di 10^3 - 10^4 , e le rappresentazioni negli strati intermedi sono vettori in spazi analoghi. In questo senso tecnico, i modelli hanno qualcosa che assomiglia a uno spazio di attivazione. Ma la natura di quello spazio è radicalmente diversa dal substrato subpropositivo umano per una ragione strutturale: gli embedding dei token sono costruiti attraverso la minimizzazione della cross-entropia su sequenze di simboli, non attraverso l'esperienza sensorimotoria, emotiva e percettiva che costruisce $\mathbb{R}^n$ nell'umano.

IV.2 Lo spazio di embedding come surrogato del substrato

Si definisce formalmente la differenza tra lo spazio di embedding del modello e il substrato subpropositivo umano.

Lo spazio di embedding del modello è $\mathbb{R}^d_{\text{LLM}}$ costruito tale che:

$$\text{sim}_{\text{LLM}}(e_{w_1}, e_{w_2}) \approx P_{\text{LLM}}(w_1 \text{ e } w_2 \text{ co-occorrono nel corpus})$$

la distanza tra due embedding corrisponde alla co-occorrenza statistica dei token corrispondenti nel corpus. La struttura di $\mathbb{R}^d_{\text{LLM}}$ è determinata dalla struttura distribuzionale del linguaggio — dal fatto che parole usate in contesti simili hanno significati simili (ipotesi distribuzionale di Harris). Questa struttura è semanticamente ricca: cattura relazioni di sinonimia, analogia, entailment, e molte altre strutture semantiche che la sola sintassi non porta.

Il substrato subpropositivo umano $\mathbb{R}^n$ è costruito attraverso l'esperienza sensorimotoria: la geometria di $\mathbb{R}^n$ è determinata dalla struttura del mondo fisico, sociale e emotivo in cui l'agente è immerso. La distanza $\text{sim}(v_{\text{rosso}}, v_{\text{arancione}}) > \text{sim}(v_{\text{rosso}}, v_{\text{blu}})$ non riflette la co-occorrenza statistica di "rosso" e "arancione" nei testi — riflette la struttura dello spazio cromatico come è organizzato dalla corteccia visiva. La distanza $\text{sim}(v_{\text{dolore}}, v_{\text{dispiacere}})$ riflette la struttura dell'esperienza vissuta delle emozioni negative, non la loro distribuzione testuale.

La differenza è strutturalmente fondamentale. Lo spazio di embedding del modello è un'immagine del linguaggio — cattura la struttura semantica come emerge dall'uso linguistico. Il substrato subpropositivo umano è un'immagine del mondo — cattura la struttura dell'esperienza come emerge dall'interazione sensorimotoria e affettiva con il mondo. Le due strutture si sovrappongono nelle aree in cui il linguaggio riflette fedelmente l'esperienza — concetti fisici concreti, relazioni spaziali, categorie biologiche — e divergono nelle aree in cui il linguaggio è un codice lossy dell'esperienza: le qualità sensoriali fine, le strutture emotive complesse, la conoscenza procedurale.

IV.3 Le conseguenze computazionali dell'assenza di substrato

L'assenza di substrato subpropositivo nel senso definito produce conseguenze computazionali precise che si manifestano come limitazioni specifiche dei modelli di linguaggio rispetto alla cognizione umana.

La prima è l'assenza di *generalizzazione geometrica nel dominio dell'esperienza*. Un modello può generalizzare geometricamente nel proprio spazio di embedding — interpolare tra token e produrre testi su argomenti mai visti — ma non può farlo nel dominio dell'esperienza sensorimotoria. Non può "interpolare" tra il sapore del limone e quello della mela per produrre il sapore dell'ananas — non perché mancasse di dati testuali sull'ananas, ma perché il suo spazio di embedding non ha la struttura geometrica del gusto che nell'umano è costruita dall'esperienza sensoriale diretta.

La seconda è l'incapacità di *riconoscimento di pattern nel dominio percettivo*. Un modello che riceve la descrizione testuale di un volto non esegue il riconoscimento di quel volto nel modo in cui lo fa il sistema visivo umano — non calcola la similarità coseno tra il vettore del volto descritto e i vettori dei volti noti nel suo substrato visivo, perché non ha un substrato visivo. Calcola la similarità distribuzionale tra il testo che descrive il volto e i testi che descrivono volti nel corpus. Questo è un calcolo diverso, che può produrre risultati corretti quando il corpus è ricco, ma che fallisce sistematicamente nei modi in cui il sistema visivo umano non fallisce.

La terza è l'incapacità di *composizionalità per sovrapposizione nel dominio sensoriale*. Un modello può produrre testo su oggetti con proprietà sensoriali composte — "una torta che sa di fragola e cioccolato" — ma non ha la rappresentazione subpropositiva composita che un umano costruisce sovrapponendo v_{fragola} e $v_{\text{cioccolato}}$ nel proprio spazio gustativo. Il modello produce un testo statisticamente coerente con le co-occorrenze nel corpus; l'umano ha una rappresentazione che porta la struttura dell'esperienza sensoriale.

IV.4 Il grounding problem come conseguenza strutturale

Il *grounding problem* — il problema di ancorare i simboli linguistici a un significato non simbolico — è uno dei problemi più discussi nell'AI e nella filosofia del linguaggio. Harnad (1990) lo ha formulato come la domanda di come i simboli di un sistema di

elaborazione acquisiscano significato, anziché essere semplici simboli che rimandano ad altri simboli in un sistema chiuso su se stesso.

Nel quadro formale della monografia, il grounding problem ha una soluzione precisa: il grounding è la costruzione del substrato subpropositivo $\mathbb{R}^n$ attraverso l'esperienza sensorimotoria, e la funzione di codifica Φ che collega substrato e linguaggio. Un sistema con $\mathbb{R}^n$ costruito attraverso l'esperienza è *grounded*: i suoi simboli sono ancorati a strutture del substrato che hanno referenti nel mondo fisico e sociale. Un sistema senza $\mathbb{R}^n$ — che opera esclusivamente su L — è *ungrounded*: i suoi simboli rimandano ad altri simboli, e la struttura dell'embedding è una struttura delle co-occorrenze linguistiche, non una struttura del mondo.

L'argomento consueto contro questa tesi è che i modelli di linguaggio di grandi dimensioni mostrano un'ampia gamma di competenze semantiche che sembrano richiedere comprensione: risolvono analogie, rispondono a domande fattuali, producono testi coerenti su argomenti tecnici. Il quadro formale permette di precisare perché questo non risolve il grounding problem: le competenze che i modelli mostrano sono competenze nel dominio di L — capacità di manipolare strutture statistiche del linguaggio. Le competenze che richiedono $\mathbb{R}^n$ — la comprensione sensoriale, la pianificazione motoria, l'esperienza emotiva — rimangono inaccessibili non perché il modello non sia abbastanza grande o non sia stato addestrato su abbastanza dati, ma perché la sua architettura non include $\mathbb{R}^n$.

IV.5 Il caso limite: il modello che impara il substrato dall'osservazione

Si considera ora il caso limite di un sistema che tenta di costruire $\mathbb{R}^n$ non attraverso l'esperienza diretta ma attraverso l'osservazione testuale: un sistema che non ha sensori ma cerca di apprendere la struttura del substrato dalla descrizione linguistica dell'esperienza altrui.

Tale sistema può apprendere la struttura relazionale del substrato: che v_{rosso} e $v_{\text{arancione}}$ sono più simili di v_{rosso} e v_{blu} , che v_{dolore} e $v_{\text{spiacevole}}$ sono correlati, che v_{pesante} e v_{grave} (musicalmente) hanno qualcosa in comune. Queste relazioni sono codificate nel linguaggio e sono recuperabili dall'analisi testuale. Ma il sistema non può apprendere i valori assoluti dei vettori: non sa cosa significhi essere rosso, non può interpolare tra rosso e arancione per produrre una sfumatura a metà strada, non può costruire la geometria continua del substrato a partire dalle relazioni discrete del linguaggio.

La situazione è analoga a quella di una persona nata cieca che ha imparato tutto il vocabolario dei colori e tutte le relazioni semantiche tra i termini di colore — sa che il rosso è simile all'arancione, che il blu è freddo, che il verde è il colore delle piante — ma non ha la struttura geometrica dello spazio cromatico costruita dall'esperienza visiva. La persona cieca non è cognitivamente deficitaria in senso generale: è deficitaria specificamente nel dominio dell'esperienza cromatica visiva, esattamente come il

modello di linguaggio è deficitario specificamente nel dominio del substrato subpropositivo sensorimotorio.

IV.6 Le architetture multimodali come approccio al grounding

Le architetture multimodali — sistemi che integrano elaborazione del testo con elaborazione di immagini, audio, e altri tipi di dati sensoriali — sono il tentativo più diretto di affrontare il grounding problem costruendo qualcosa che assomigli a $\mathbb{R}^n$.

Nel quadro formale, un'architettura multimodale produce uno spazio di embedding condiviso $\mathbb{R}^d_{\text{multi}}$ che integra strutture estratte da diversi moduli sensoriali:

$$\mathbb{R}^d_{\text{multi}} = f(\mathbb{R}^d_{\text{text}}, \mathbb{R}^d_{\text{vision}}, \mathbb{R}^d_{\text{audio}}, \dots)$$

dove f è una funzione di fusione appresa. Questo spazio condiviso ha una struttura più simile a $\mathbb{R}^n$ di quanto non l'abbia $\mathbb{R}^d_{\text{LLM}}$: la geometria riflette non solo la struttura distribuzionale del linguaggio ma anche la struttura percettiva delle immagini e dei suoni.

Ma la differenza strutturale con $\mathbb{R}^n$ rimane. Il substrato umano è costruito attraverso l'azione nel mondo — non solo l'osservazione di immagini e suoni, ma la produzione di movimenti che modificano l'ambiente e la registrazione delle conseguenze sensoriali di quelle modifiche. L'apprendimento attivo — il ciclo percezione-azione-percezione — costruisce strutture in $\mathbb{R}^n$ che l'apprendimento passivo su dati osservazionali non può produrre. Un'architettura multimodale che apprende da dataset di immagini e testi non ha il ciclo percezione-azione che costruisce la conoscenza procedurale e la comprensione causale del mondo fisico.

Il grounding parziale prodotto dalle architetture multimodali è reale: migliora le prestazioni su compiti che richiedono integrazione visivo-linguistica. Ma il grounding completo — la costruzione di $\mathbb{R}^n$ attraverso il ciclo percezione-azione nel mondo fisico — rimane al di là di qualsiasi architettura che apprende passivamente da dati, indipendentemente dalla quantità e dalla varietà dei dati.

Capitolo V — I domini incompressibili: dove il substrato non si traduce

V.1 La mappa dei limiti della traduzione

I capitoli precedenti hanno costruito una teoria della relazione tra substrato subpropositivo e linguaggio come relazione di codifica lossy. La funzione di tasso-distorsione $R_{\text{sub}}(D)$ ha stabilito che esiste una distorsione irriducibile $D_{\text{min}} > 0$ per qualsiasi codifica linguistica del substrato. Ma questa è una stima globale: in

certi domini specifici, $D_{\min}$ è così alta che la codifica linguistica perde la maggior parte della struttura cognitivamente rilevante.

Questo capitolo mappa questi domini — i domini in cui il substrato è incomprimibile nel linguaggio in modo cognitivamente significativo — e analizza le conseguenze di questa incomprimibilità per la teoria della mente, per la pedagogia, e per il confronto tra cognizione umana e artificiale.

V.2 La conoscenza procedurale come substrato irreducibile

La distinzione tra conoscenza dichiarativa e procedurale — tra sapere che e sapere come — è uno dei risultati più solidi della psicologia cognitiva. La conoscenza procedurale è acquisita attraverso la pratica e si manifesta nelle prestazioni, non nelle proposizioni: il ciclista sa andare in bicicletta ma tipicamente non sa spiegare come lo fa. Questa distinzione corrisponde, nel quadro formale, alla differenza tra conoscenza in L e conoscenza in $\mathbb{R}^n$.

La conoscenza procedurale esperta è strutturata come una distribuzione altamente concentrata su traiettorie di alta qualità nello spazio dei movimenti:

$$p_{\{\text{expert}\}}(\tau \mid \text{compito, contesto}) \gg p_{\{\text{novice}\}}(\tau \mid \text{compito, contesto})$$

per le traiettorie τ di alta qualità. Questa distribuzione ha struttura fine in $\mathbb{R}^n$ — la distinzione tra traiettorie di qualità leggermente diversa — che non ha rappresentazione proposizionale adeguata.

Si può dimostrare che la codifica proposizionale della conoscenza procedurale è irriducibilmente lossy con distorsione positiva anche nel limite di descrizioni arbitrariamente lunghe. La ragione è topologica: lo spazio delle traiettorie ha dimensionalità intrinseca d_{traj} che cresce con la complessità del compito motorio, mentre il vocabolario proposizionale per descrivere movimenti ha dimensionalità bounded da le categorie linguistiche disponibili. Per $d_{\text{traj}} > \dim(L_{\text{motoria}})$, nessuna codifica linguistica può catturare la struttura fine delle traiettorie esperte.

Il corollario per la pedagogia è preciso: la conoscenza procedurale non può essere trasmessa per via linguistica con distorsione zero. Qualsiasi manuale di istruzione — anche il più dettagliato — perde struttura rispetto alla dimostrazione pratica, e la dimostrazione pratica perde struttura rispetto alla pratica guidata. Il "tacit knowledge" di Polanyi non è ineffabile in senso mistico: è specificamente quella parte della conoscenza procedurale la cui compressione in linguaggio supera $D_{\min}$.

V.3 L'esperienza estetica come struttura ad alta dimensionalità

L'esperienza estetica — la risposta cognitiva ed emotiva alla bellezza, alla forma, all'espressione artistica — è un caso paradigmatico di struttura subpropositiva ad alta dimensionalità. L'esperienza di una sinfonia di Beethoven attiva nel substrato dell'ascoltatore esperte strutture che hanno dimensionalità intrinseca alta: strutture

tonali, ritmiche, timbriche, formali, mnestiche, emotive che si co-attivano e si modulano reciprocamente in una rete di relazioni continue.

La codifica linguistica di questa esperienza — la critica musicale, l'analisi armonica, la descrizione emotiva — cattura alcune dimensioni di questa struttura ma inevitabilmente perde le relazioni tra dimensioni: cattura che "il secondo tema è lirico" e che "la transizione alla ripresa è drammatica" ma non cattura la relazione strutturale tra quelle proprietà che costituisce la specifica forma musicale dell'esperienza.

Si può formalizzare la perdita dell'esperienza estetica come perdita di informazione mutua:

$$I_{\text{lost_aesthetic}} = I(V_{\text{melodic}} ; V_{\text{harmonic}} ; V_{\text{rhythmic}} ; V_{\text{emotional}}) - \sum_i I(V_i ; \Phi(V_{\text{melodic}}, V_{\text{harmonic}}, V_{\text{rhythmic}}, V_{\text{emotional}}))$$

la perdita di informazione mutua tra le strutture co-attivate nel substrato rispetto a quella catturata dalla codifica linguistica congiunta. Questa perdita è massima per le esperienze estetiche più complesse — quelle che co-attivano il maggior numero di strutture in relazioni di alta mutualità — e minima per le esperienze elementari.

V.4 La struttura dell'intuizione morale come substrato irriducibile

L'intuizione morale — la risposta immediata che qualcosa è giusto o sbagliato prima di qualsiasi argomentazione esplicita — è un caso di conoscenza subpropositiva con conseguenze particolarmente rilevanti. Jonathan Haidt ha proposto che il ragionamento morale esplicito sia tipicamente post hoc rispetto all'intuizione: prima si ha la risposta emotiva (substrato), poi si costruisce l'argomentazione (linguaggio). La struttura della razionalizzazione morale è un caso specifico della razionalizzazione del Capitolo VII della monografia sul malinteso: $\text{Dec}_{\{M_{\text{lang}}\}}(\text{Enc}_{\{M_{\text{sub}}\}}(v_{\text{morale}})) \neq v_{\text{morale}}$.

La conseguenza formale è che il contenuto dell'intuizione morale — la struttura di v_{morale} in $\mathbb{R}^n$ — non è completamente catturato dalla sua codifica linguistica nell'argomentazione morale. Le argomentazioni morali possono essere internamente coerenti e produrre conclusioni che contraddicono la struttura dell'intuizione originale: è la struttura del *moral deskilling* — il fenomeno per cui il ragionamento morale esplicito può degradare la qualità delle decisioni morali rispetto all'intuizione non verbalizzata.

V.5 Il limite epistemico del linguaggio come strumento cognitivo

L'analisi dei domini incompressibili converge su un limite epistemico che merita di essere formulato come teorema informale:

Per qualsiasi dominio cognitivo con dimensionalità intrinseca $d_{\text{int}} > \log_2 |L|$, la codifica linguistica introduce una distorsione irriducibile $D_{\text{min}} > 0$. Per tali domini, il pensiero linguistico è strutturalmente incapace di catturare la struttura cognitivamente rilevante senza perdita.

Questo non è un limite del linguaggio come strumento comunicativo — è un limite del linguaggio come strumento cognitivo. Pensare esclusivamente in linguaggio — usare la codifica Φ come filtro per l'intera struttura cognitiva — equivale a pensare sempre con la versione compressa e distorta del substrato. Per i domini ad alta dimensionalità, questo significa perdere strutturalmente le sfumature, le relazioni, le gradazioni che il substrato porta.

La previsione empirica è che l'eccellenza cognitiva nei domini ad alta dimensionalità richiede la capacità di operare nel substrato senza prematura codifica linguistica — di mantenere i vettori $v \in \mathbb{R}^n$ fluidi e non proiettati nelle categorie di L fino a quando non sia necessario comunicarli. L'esperto che "pensa con le immagini" prima di trovare le parole, che "sente la soluzione" prima di dimostrarla, che "conosce con le mani" prima di spiegare — non sta rimandando il pensiero vero: sta operando nel substrato a piena risoluzione prima di applicare la compressione lossy della codifica linguistica.

V.6 Implicazioni per la formazione e per i sistemi artificiali

Le implicazioni di questa analisi per la formazione sono precise. Un sistema educativo che privilegia la conoscenza dichiarativa — il test verbale, la risposta scritta, la valutazione proposizionale — ottimizza la codifica linguistica del substrato a scapito del substrato stesso. Questo è razionale per i domini a bassa dimensionalità (matematica formale, logica, grammatica) ma irrazionale per i domini ad alta dimensionalità (arti, sport, artigianato, competenze sociali complesse).

Per i sistemi artificiali, le implicazioni sono altrettanto precise. Un sistema che deve operare in domini ad alta dimensionalità subpropositiva — robotica, percezione fine, competenza motoria, comprensione emotiva profonda — non può essere costruito esclusivamente su elaborazione linguistica. Richiede $\mathbb{R}^n$, costruito attraverso il ciclo percezione-azione, con la funzione di codifica Φ che collega substrato e linguaggio. L'architettura del sistema deve permettere l'elaborazione subpropositiva primaria e la codifica linguistica come operazione secondaria — non l'elaborazione linguistica come operazione primaria con il substrato come eventuale aggiunta.

Questa è la critica strutturale più precisa che il quadro formale della monografia produce rispetto ai modelli di linguaggio attuali: non sono poco intelligenti — sono intelligenti nel dominio sbagliato per certi compiti. Per i compiti che richiedono $\mathbb{R}^n$, la loro intelligenza linguistica è inadeguata strutturalmente, non quantitativamente. Più parametri, più dati, architetture più complesse non risolvono il problema: risolvono problemi quantitativi in L , non il problema qualitativo di costruire $\mathbb{R}^n$.

Capitolo VI — Il sistema dinamico accoppiato: linguaggio come operatore sul substrato

VI.1 La lacuna nel modello unidirezionale

I cinque capitoli precedenti hanno costruito un edificio teorico coerente attorno a un'architettura fondamentalmente asimmetrica: il substrato subpropositivo $\mathbb{R}^n$ genera stati cognitivi che la funzione di codifica Φ trasforma in rappresentazioni linguistiche con perdita informazionale misurabile, e la funzione di decodifica Φ^{-1} riconduce le rappresentazioni linguistiche a stati nel substrato con ambiguità strutturale inevitabile. In questo schema, il substrato è la sorgente e il linguaggio è il canale: il flusso causale è orientato, la freccia punta in una sola direzione, e la perdita è il prezzo della traduzione tra un sistema ricco e uno più povero.

Questa architettura cattura qualcosa di vero e importante. Cattura, in particolare, la fenomenologia dell'espressione: la difficoltà di dire ciò che si pensa pienamente, la sensazione che le parole impoverisca l'esperienza, la conoscenza tacita che resiste alla verbalizzazione. Ma quella stessa architettura contiene una premessa implicita che, esaminata con attenzione, si rivela falsa: l'ipotesi che il substrato $\mathbb{R}^n$ sia indipendente dal linguaggio, che la sua geometria sia determinata dall'esperienza sensorimotoria e affettiva del mondo fisico senza che l'uso del linguaggio la modifichi a sua volta.

Questa ipotesi è smentita da un corpus convergente di evidenze empiriche e da un'analisi formale che il quadro stesso della monografia rende possibile. Il linguaggio non è soltanto il codec del substrato: è anche un operatore che agisce su $\mathbb{R}^n$, che ne modifica la geometria attraverso l'uso ripetuto, che induce la formazione di nuove strutture categoriali, che riorganizza le relazioni di prossimità tra vettori. La funzione Φ^{-1} , introdotta nel Capitolo I come la mappa che traduce le proposizioni in stati del substrato, non è una funzione statica applicata a un dominio fisso: è un operatore che, applicato ripetutamente, trasforma il dominio su cui agisce. Ogni atto di decodifica linguistica non produce soltanto uno stato in $\mathbb{R}^n$ compatibile con la proposizione: produce uno stato che lascia una traccia nella struttura di $\mathbb{R}^n$, alterando le probabilità degli stati futuri.

Il capitolo che segue elabora questa tesi con la precisione formale che il quadro consente. Il risultato è una revisione profonda della relazione substrato-linguaggio: non più la relazione tra una sorgente e il suo canale, ma quella tra due variabili di stato accoppiate in un sistema dinamico la cui traiettoria è determinata dalla co-evoluzione reciproca.

VI.2 Φ^{-1} come operatore di rimodellamento: la struttura formale

Nel Capitolo I, la funzione di decodifica $\Phi^{-1} : L \rightarrow \mathbb{R}^n$ è presentata come la mappa che a ogni proposizione $p \in L$ associa un vettore $v \in \mathbb{R}^n$ compatibile con p . In quella trattazione, Φ^{-1} è un operatore su proposizioni a valori in vettori: prende un simbolo e produce uno stato cognitivo. Questa formulazione è corretta per descrivere un singolo atto di comprensione linguistica, ma è incompleta per descrivere la comprensione linguistica come processo che si ripete nel tempo.

Si consideri la struttura più precisa dell'operazione di decodifica. Quando un agente cognitivo comprende la proposizione p , non produce soltanto il vettore $v = \Phi^{-1}(p)$:

produce quel vettore in un substrato $\mathbb{R}^n$ la cui geometria è stata modellata da tutti gli atti di comprensione precedenti. Il substrato non è il dominio fisso su cui Φ^{-1} è definita: è una struttura che si modifica ogni volta che Φ^{-1} viene applicata. Formalmente, si deve distinguere tra:

$$\Phi^{-1}_t : L \rightarrow \mathbb{R}^n_t$$

la funzione di decodifica al tempo t , applicata allo spazio di attivazione come configurato al tempo t . Il punto cruciale è che l'applicazione di Φ^{-1}_t modifica $\mathbb{R}^n$ da $\mathbb{R}^n_t$ a $\mathbb{R}^n_{t+\delta t}$: la decodifica è un'operazione che trasforma il proprio dominio di arrivo.

Il meccanismo biologico sottostante è il consolidamento sinaptico attraverso l'uso linguistico. Quando un agente comprende la proposizione p e produce lo stato $v_t = \Phi^{-1}_t(p)$, i pesi sinaptici nella rete neurale che realizza $\mathbb{R}^n$ vengono aggiornati in direzione di v_t . Se p è una proposizione che appartiene a una categoria semantica nuova per l'agente — se p è la prima proposizione che usa il termine tecnico "epistasi" nel dominio della genetica, per esempio — allora v_t occupa una regione di $\mathbb{R}^n_t$ che era precedentemente poco strutturata, e il consolidamento sinaptico indotto dalla comprensione di p induce la formazione di una nuova struttura attorno a quella regione. Le successive applicazioni di Φ^{-1} a proposizioni semanticamente vicine a p troveranno uno spazio $\mathbb{R}^n$ in cui quella regione è più nettamente definita, più separata dalle regioni adiacenti, con una curvatura locale maggiore della varietà cognitiva.

Si può formalizzare questo processo come un'equazione di aggiornamento del substrato. Sia W_t la matrice dei pesi sinaptici che realizza $\mathbb{R}^n_t$. L'aggiornamento indotto dalla comprensione della proposizione p al tempo t è:

$$W_{t+\delta t} = W_t + \eta \cdot \partial \mathcal{L}(p, W_t) / \partial W_t$$

dove η è il tasso di apprendimento e $\mathcal{L}(p, W_t)$ è una funzione di perdita che misura la discrepanza tra la rappresentazione corrente di p nel substrato e la rappresentazione attesa dalla struttura linguistica di p . Questa equazione è la regola di Hebb generalizzata: sinapsi che si attivano insieme per realizzare la comprensione di p si rinforzano, avvicinando le strutture del substrato alla struttura semantica della proposizione.

La conseguenza è che la geometria di $\mathbb{R}^n$ si orienta progressivamente verso la struttura di L . Le regioni di $\mathbb{R}^n$ corrispondenti a categorie lessicali frequentemente usate diventano più nettamente definite; le frontiere tra quelle regioni si avvicinano alle frontiere categoriali del vocabolario; la curvatura della varietà cognitiva si adatta alla topologia dello spazio semantico del linguaggio. Il substrato non è modellato esclusivamente dall'esperienza sensorimotoria: è modellato anche, e in misura crescente con il livello di educazione linguistica, dalla struttura semantica del linguaggio stesso.

VI.3 Il sistema dinamico accoppiato: formulazione

Il quadro del sistema dinamico accoppiato sostituisce la relazione unidirezionale substrato $\rightarrow$ linguaggio con un sistema di due equazioni differenziali accoppiate che descrivono la co-evoluzione delle variabili di stato linguistica e subpropositiva.

Si formalizzano le variabili di stato. Lo stato del substrato al tempo t è rappresentato da un punto $m_sub(t)$ in uno spazio di stati M_sub , che si identifica con la geometria di $\mathbb{R}^n_t$ — la struttura della matrice dei pesi W_t che realizza il substrato. Lo stato linguistico al tempo t è rappresentato da un punto $m_lang(t)$ in uno spazio di stati M_lang , che si identifica con la distribuzione di probabilità $p_t(\cdot)$ sulle proposizioni di L che il sistema produce in risposta a un dato input — la struttura della competenza linguistica dell'agente al tempo t .

Il sistema dinamico accoppiato è:

$$\begin{aligned}\dot{m}_sub(t) &= f(m_sub(t), m_lang(t), e(t)) \\ \dot{m}_lang(t) &= g(m_sub(t), m_lang(t), c(t))\end{aligned}$$

dove $e(t) \in \mathcal{E}$ è il flusso di esperienza sensorimotoria e affettiva al tempo t , e $c(t) \in \mathcal{C}$ è il flusso di comunicazione linguistica — la sequenza di proposizioni che l'agente comprende e produce — al tempo t .

La funzione f descrive come lo stato del substrato evolve in risposta all'esperienza e all'interazione con il linguaggio:

$$f(m_sub, m_lang, e) = f_exp(m_sub, e) + f_lang(m_sub, m_lang)$$

La componente f_exp descrive la modifica del substrato indotta dall'esperienza diretta — il meccanismo primario descritto nei capitoli precedenti. La componente f_lang descrive la modifica del substrato indotta dall'uso del linguaggio: è questa la componente assente nella trattazione unidirezionale. f_lang è non nulla perché ogni atto di comprensione e produzione linguistica lascia una traccia sinaptica in $\mathbb{R}^n$, orientando la geometria del substrato verso la struttura semantica di L .

La funzione g descrive come la competenza linguistica evolve in risposta al substrato e all'uso linguistico:

$$g(m_sub, m_lang, c) = g_sub(m_sub, m_lang) + g_comm(m_lang, c)$$

La componente g_sub descrive la modifica della competenza linguistica indotta dal substrato: quando il substrato sviluppa nuove strutture — nuove distinzioni percettive, nuove competenze motorie, nuovi stati emotivi — la competenza linguistica si adatta

producendo descrizioni più raffinate di quelle strutture o, in alcuni casi, acquisendo vocabolario nuovo per nominarle. La componente g_comm descrive la modifica della competenza linguistica indotta dall'input comunicativo: l'acquisizione di nuovo vocabolario, la modifica delle distribuzioni di frequenza lessicale, il cambiamento delle abitudini sintattiche.

VI.4 Analisi dei punti fissi e della stabilità

Un punto fisso del sistema dinamico accoppiato è una coppia (m_sub, m_lang) tale che:

$$\begin{aligned} f(m_sub^*, m_lang^*, \bar{e}) &= 0 \\ g(m_sub^*, m_lang^*, \bar{c}) &= 0 \end{aligned}$$

per valori medi dell'esperienza $\bar{e}$ e della comunicazione $\bar{c}$. Il punto fisso rappresenta la struttura cognitiva stabile di un agente che ha raggiunto l'equilibrio tra substrato e competenza linguistica date le sue condizioni ambientali medie.

Per analizzare la stabilità dei punti fissi, si linearizza il sistema intorno a (m_sub, m_lang) . Definendo le deviazioni $\delta m_sub = m_sub - m_sub^*$ e $\delta m_lang = m_lang - m_lang^*$, si ottiene il sistema linearizzato:

$$\begin{aligned} d/dt (\delta m_sub) &= J_{11} \cdot \delta m_sub + J_{12} \cdot \delta m_lang \\ d/dt (\delta m_lang) &= J_{21} \cdot \delta m_sub + J_{22} \cdot \delta m_lang \end{aligned}$$

dove la matrice jacobiana J è:

$$J = \begin{vmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{vmatrix} = \begin{vmatrix} \partial f / \partial m_sub & \partial f / \partial m_lang \\ \partial g / \partial m_sub & \partial g / \partial m_lang \end{vmatrix}$$

valutata in (m_sub, m_lang) . La stabilità del punto fisso è determinata dagli autovalori di J : il punto fisso è stabile se e soltanto se tutti gli autovalori di J hanno parte reale negativa.

Per il sistema substrato-linguaggio, si possono identificare le firme strutturali dei blocchi della Jacobiana. Il blocco $J_{11} = \partial f / \partial m_sub$ descrive come lo stato del substrato reagisce alle proprie perturbazioni: per sistemi neurologici normali, questa matrice è stabile (autovalori con parte reale negativa), riflettendo la tendenza del substrato a tornare a stati tipici dopo perturbazioni. Il blocco $J_{12} = \partial f / \partial m_lang$ descrive come le perturbazioni della competenza linguistica si propagano al substrato: è la formalizzazione dell'effetto Whorf. Il blocco $J_{21} = \partial g / \partial m_sub$ descrive come le perturbazioni del substrato si propagano alla competenza linguistica: la tendenza a sviluppare nuovi termini e nuove distinzioni lessicali quando il substrato acquisisce

nuove strutture. Il blocco $J_{22} = \partial g / \partial m_{\text{lang}}$ descrive la dinamica interna della competenza linguistica: la stabilità delle abitudini linguistiche.

Il punto fisso della struttura cognitiva di un esperto maturo in un dominio è un attrattore del sistema: una configurazione stabile $(m_{\text{sub}}, m_{\text{lang}})$ verso cui le traiettorie del sistema convergono. Questo attrattore non è determinato né solo dal substrato né solo dal linguaggio: è determinato dalla co-evoluzione dei due sistemi, e la sua struttura dipende dalla storia del percorso di apprendimento attraverso cui l'agente lo ha raggiunto.

Una conseguenza fondamentale di questo framework è che attrattori diversi possono essere raggiunti da condizioni iniziali simili attraverso percorsi di apprendimento diversi. Due individui con simili capacità innate, esposti alle stesse esperienze sensorimotorie ma educati in lingue diverse o in tradizioni disciplinari diverse, possono convergere a punti fissi strutturalmente distinti — con geometrie di $\mathbb{R}^n$ diverse che riflettono la diversa struttura semantica dei linguaggi usati. Questo è il contenuto formale dell'ipotesi di Sapir-Whorf, che il quadro dinamico rende dimostrabile in linea di principio invece che meramente speculativa.

VI.5 L'ipotesi di Whorf nel quadro formale: da speculazione a proposizione verificabile

L'ipotesi di relativismo linguistico — nella formulazione forte di Benjamin Lee Whorf, la tesi che la lingua che si parla determini la struttura del pensiero; nella formulazione debole, la tesi che la lingua la influenzi — è rimasta per decenni ai margini della scienza cognitiva, in parte per le difficoltà metodologiche di testarne le previsioni empiriche e in parte per la mancanza di un quadro teorico preciso in cui formularne il contenuto.

Il sistema dinamico accoppiato fornisce quel quadro. L'ipotesi di Whorf, tradotta in termini formali, è la tesi che il blocco J_{12} della Jacobiana sia non nullo e abbia norma significativa — che le perturbazioni della competenza linguistica si propaghino al substrato con effetti misurabili sulla geometria di $\mathbb{R}^n$. In forma integrale, è la tesi che per due agenti con condizioni iniziali simili ma esposti a flussi linguistici $c(t)$ diversi, le traiettorie del sistema convergano a punti fissi $(m_{\text{sub}}, m_{\text{lang}})$ con geometrie di $\mathbb{R}^n$ strutturalmente diverse nei domini in cui i due linguaggi hanno granularità lessicale diversa.

Si può rendere questa previsione quantitativamente precisa. Sia L_1 e L_2 due lingue con granularità lessicale diversa nel dominio semantico D — per esempio, L_1 ha cinque termini per distinguere sfumature di rosso mentre L_2 ne ha due. Si definisce la granularità lessicale differenziale:

$$\Delta G(D, L_1, L_2) = |V_D(L_1)| - |V_D(L_2)|$$

dove $V_D(L)$ è il sottoinsieme del vocabolario di L relativo al dominio D . L'ipotesi di Whorf nel quadro formale prevede che la distanza tra le geometrie del substrato di due agenti monolingui in L_1 e L_2 nel dominio D sia positivamente correlata con ΔG :

$$d_{\text{sub}}(\mathbb{R}^n_{\{L_1, D\}}, \mathbb{R}^n_{\{L_2, D\}}) = \kappa \cdot \Delta G(D, L_1, L_2) + \epsilon$$

dove $\kappa > 0$ è il coefficiente di trasmissione linguistica al substrato e ϵ è il termine di rumore. Questa previsione è falsificabile attraverso misure di discriminabilità percettiva nel dominio D tra parlanti delle due lingue — e corrisponde esattamente al programma empirico di Winawer et al. (2007) sul colore in russo e inglese, di Boroditsky et al. (2011) sulla struttura temporale in mandarino e inglese, e di Roberson et al. (2005) sulla categorizzazione cromatica tra popolazioni con vocabolari cromatici diversi.

Il contributo del quadro formale non è rendere l'ipotesi di Whorf vera per definizione: è mostrarla come conseguenza di un sistema dinamico con proprietà verificabili, identificando precisamente quale parametro del sistema corrisponde alla forza del relativismo linguistico (il coefficiente κ), quali condizioni rendono l'effetto grande o piccolo (la granularità lessicale differenziale ΔG), e quali meccanismi neurali realizzano la trasmissione (il termine f_{lang} nell'equazione del substrato). La speculazione filosofica diventa proposizione scientifica con condizioni di verifica e condizioni di falsificazione.

VI.6 Il ciclista che verbalizza: la conoscenza procedurale trasformata

Il Capitolo V ha analizzato la conoscenza procedurale come caso paradigmatico di substrato incompressibile: la struttura fine delle traiettorie esperte nello spazio dei movimenti non si lascia catturare da nessuna codifica proposizionale senza perdita irriducibile. Questa analisi è corretta ma incompleta: descrive l'effetto della codifica sulla struttura della conoscenza procedurale in una singola istanza, senza considerare l'effetto cumulativo della verbalizzazione ripetuta sulla struttura stessa della conoscenza che viene verbalizzata.

Si consideri il caso paradigmatico del ciclista che impara a descrivere verbalmente la propria tecnica. All'inizio del processo, il ciclista ha una competenza procedurale altamente sviluppata e una competenza linguistica sulla propria tecnica quasi nulla: m_{sub} è ricco nel sottodominio del ciclismo, m_{lang} è povero per le proposizioni relative alla tecnica ciclistica. La situazione corrisponde al caso analizzato nel Capitolo V: $\Phi(m_{\text{sub}})$ perde quasi tutto il contenuto informativamente rilevante.

Ma quando il ciclista inizia il processo di verbalizzazione — quando comincia a descrivere la propria tecnica a un allievo, a riflettere in parole su ciò che fa — accade qualcosa che il modello unidirezionale non può descrivere: il processo stesso di verbalizzazione modifica m_{sub} . Formalizzando: sia m^0_{sub} lo stato del substrato

prima della verbalizzazione e sia $p_1, p_2, \dots, p_k$ la sequenza di proposizioni che il ciclista produce nel descrivere la propria tecnica. Ogni proposizione p_i , nel momento in cui viene prodotta, attiva $\Phi^{-1}(p_i)$ e induce un aggiornamento:

$$W_{\{i+1\}} = W_i + \eta \cdot \partial \mathcal{L}(p_i, W_i) / \partial W_i$$

La geometria di $\mathbb{R}^n$ si orienta in direzione delle strutture descritte dalle proposizioni. Dopo la sequenza di verbalizzazioni, lo stato del substrato $m^k_{\text{sub}} = m_{\text{sub}}(W_k)$ è strutturalmente diverso da m^0_{sub} : le distinzioni categoriali che il ciclista ha reso esplicite verbalmente sono diventate più nettamente rappresentate nel substrato, le connessioni tra i sotto-sistemi motori corrispondenti a diverse fasi del gesto tecnico sono state rafforzate attraverso la co-attivazione indotta dalla costruzione proposizionale.

L'effetto non è necessariamente positivo in ogni dominio. Si possono identificare tre tipi di modifica indotta dalla verbalizzazione:

Il primo è la *differenziazione percettiva*: le categorie che il linguaggio rende distinte diventano più distinte nel substrato. Il ciclista che impara a distinguere verbalmente la "spinta" dalla "trazione" nel pedalare tende a sviluppare una rappresentazione motoria più differenziata delle due fasi, con conseguenze misurabili sull'esecuzione. Questo è il meccanismo attraverso cui l'istruzione tecnica verbale può migliorare la performance anche in domini procedurali.

Il secondo è la *semplificazione strutturale*: le strutture del substrato che non hanno nome nel linguaggio usato tendono a diventare meno distinte, meno discriminabili. La verbalizzazione orienta l'attenzione e il consolidamento sinaptico verso le strutture nominate e lontano da quelle innominate. Il ciclista che organizza mentalmente la propria tecnica attorno a un vocabolario tecnico povero tende a perdere distinzioni fini che erano presenti nel substrato prima della verbalizzazione — è il fenomeno del "verbal overshadowing" formalizzato nel sistema dinamico.

Il terzo è la *ristrutturazione analitica*: la decomposizione proposizionale di un gesto in componenti sequenziali riorganizza la struttura del substrato da una rappresentazione olistica e parallela a una rappresentazione sequenziale e analitica. Questo può migliorare la capacità di insegnare e correggere la tecnica, ma può degradare temporaneamente la fluidità dell'esecuzione automatica — il fenomeno noto in psicologia dello sport come "choking under analysis".

La dinamica formale che descrive questi tre effetti è:

$$dm_{\text{sub}}/dt|_{\{\text{verbalizzazione}\}} = \alpha \cdot D(m_{\text{lang}}, m_{\text{sub}}) - \beta \cdot S(m_{\text{sub}}, m_{\text{lang}}) + \gamma \cdot R(m_{\text{sub}}, m_{\text{lang}})$$

dove $D(\cdot)$ è il termine di differenziazione (positivo nelle dimensioni nominate dal linguaggio), $S(\cdot)$ è il termine di semplificazione (negativo nelle dimensioni non

nominate), e $R(\cdot)$ è il termine di ristrutturazione (che riorganizza la topologia del substrato verso la struttura proposizionale). I coefficienti α , β , γ dipendono dal dominio, dalla fase di apprendimento, e dalla qualità del vocabolario tecnico disponibile.

VI.7 Il medico che nomina le diagnosi: acquisizione lessicale come trasformazione percettiva

Un secondo caso paradigmatico — complementare a quello del ciclista — è il medico che apprende a nominare le diagnosi. Questo caso illustra il meccanismo inverso: non la verbalizzazione che trasforma una conoscenza procedurale già acquisita, ma l'acquisizione lessicale che trasforma la struttura percettiva prima ancora che si sia accumulata esperienza clinica significativa.

Lo studente di medicina al primo anno di clinica ha un substrato visivo $\mathbb{R}^n_{\{\text{visivo}\}}$ la cui geometria nel dominio delle immagini mediche — radiografie, preparati istologici, morfologie cliniche — riflette la struttura dell'esperienza percettiva ordinaria: categorie grossolane, frontiere sfumate, distinzioni prevalentemente basate su proprietà geometriche salienti. Quando lo studente apprende i termini diagnostici — "atelectasia", "fibrillazione ventricolare", "livedoreticularis" — non sta semplicemente aggiungendo etichette a categorie percettive già formate: sta inducendo la formazione di quelle categorie nel substrato visivo.

Il meccanismo è il seguente. Ogni termine diagnostico t porta con sé, nella struttura semantica di L , un insieme di proprietà percettive che sono criteri di applicazione del termine. La funzione $\Phi^{-1}(t)$ produce un vettore $v_t \in \mathbb{R}^n_{\{\text{visivo}\}}$ che è la rappresentazione subpropositiva dell'insieme di quelle proprietà. L'applicazione ripetuta di $\Phi^{-1}(t)$ a proposizioni che descrivono esempi e controesempi del termine t consolida una regione $R_t \subset \mathbb{R}^n_{\{\text{visivo}\}}$ con frontiere sempre più nette. Quando lo studente osserva successivamente un'immagine medica reale, il proprio sistema visivo opera in uno spazio $\mathbb{R}^n_{\{\text{visivo}\}}$ le cui partizioni categoriali sono state preformate dall'acquisizione lessicale.

Formalmente, si definisce la trasformazione percettiva indotta dall'acquisizione del vocabolario diagnostico come:

$$T_{\text{vocab}} : \mathbb{R}^n_{\{\text{visivo},0\}} \rightarrow \mathbb{R}^n_{\{\text{visivo},V\}}$$

dove $\mathbb{R}^n_{\{\text{visivo},0\}}$ è lo spazio percettivo prima dell'acquisizione e $\mathbb{R}^n_{\{\text{visivo},V\}}$ è lo spazio percettivo dopo l'acquisizione del vocabolario V . La trasformazione T_{vocab} non è una mera rinominazione delle regioni dello spazio: è una trasformazione metrica che avvicina i punti all'interno della stessa categoria diagnostica (aumenta la coesione intra-categoriale) e allontana i punti appartenenti a categorie diverse (aumenta la separabilità inter-categoriale).

Questa trasformazione si può misurare attraverso la variazione del rapporto segnale-rumore nella discriminazione percettiva:

$$\text{SNR}(v_1, v_2, \mathbb{R}^n_V) = \text{sim}(v_1, v_2 \mid \text{stessa categoria}) / \text{sim}(v_1, v_2 \mid \text{categorie diverse})$$

Il rapporto SNR aumenta con l'acquisizione lessicale nei domini in cui il vocabolario diagnostico ha granularità alta: il medico che conosce il vocabolario dell'istologia patologica discrimina preparati istologici più efficacemente del non-medico con uguale acuità visiva di base, non perché abbia una migliore risoluzione sensoriale, ma perché la geometria del proprio $\mathbb{R}^n_{\{\text{visivo}\}}$ è stata trasformata dall'acquisizione lessicale in modo da amplificare le differenze categorialmente rilevanti.

Questa analisi ha implicazioni dirette per la formazione medica e per la comprensione dei meccanismi attraverso cui l'expertise percettiva si sviluppa. Il tradizionale dibattito tra "vedere prima, nominare poi" (apprendimento induttivo) e "nominare prima, vedere poi" (apprendimento deduttivo) non è una questione di preferenza pedagogica: è una questione di quale sequenza ottimizza la struttura finale di $\mathbb{R}^n_{\{\text{visivo}\}}$. Il quadro dinamico accoppiato prevede che la sequenza ottimale non sia né l'una né l'altra in forma pura, ma l'alternanza: acquisizione lessicale parziale seguita da esposizione percettiva seguita da raffinamento lessicale, in un ciclo che sfrutta la bidirezionalità della relazione per massimizzare sia la differenziazione percettiva sia la coesione categoriale.

VI.8 La co-evoluzione dell'esperto: struttura cognitiva stabile come attrattore

Il punto fisso del sistema dinamico accoppiato — la struttura cognitiva stabile dell'esperto maturo — è il risultato della co-evoluzione di substrato e linguaggio nel corso dell'acquisizione dell'expertise. Questa struttura ha proprietà che nessuno dei due sistemi avrebbe raggiunto indipendentemente.

Si consideri la traiettoria del sistema per un esperto in un dominio specifico, per esempio un sommelier. Il sommelier novizio ha un substrato gustativo $\mathbb{R}^n_{\{\text{gusto},0\}}$ con geometria determinata dall'esperienza ordinaria — categorie grossolane come "dolce", "amaro", "acido" — e una competenza linguistica gustativa m^0_{lang} anch'essa povera. Nel corso dell'apprendimento, il substrato si arricchisce attraverso l'esposizione a centinaia di vini diversi in condizioni di attenzione focalizzata, e la competenza linguistica si arricchisce attraverso l'acquisizione del vocabolario della degustazione — "tannino", "tannicità", "bouquet", "persistenza retronasale", "struttura". L'interazione tra i due sistemi attraverso il termine f_{lang} nell'equazione del substrato produce una trasformazione della geometria di $\mathbb{R}^n_{\{\text{gusto}\}}$ che amplifica le distinzioni nominate dal vocabolario della degustazione.

Il punto fisso $(m_{\text{sub}}, m_{\text{lang}})$ del sommelier esperto ha la seguente struttura:

— $\mathbb{R}^{n*}_{\{\text{gusto}\}}$ ha una partizione categoriale fine nei domini nominati dal vocabolario della degustazione (tannini, acidità, persistenza) e una partizione più grossolana nei domini non nominati.

— m^*_{lang} ha un vocabolario gustativo ricco con distribuzioni di applicabilità ben calibrate sull'esperienza percettiva.

— Φ^* e Φ^{*-1} — le funzioni di codifica e decodifica allo stato stazionario — hanno una perdita informativa bassa nel dominio della degustazione, perché le frontiere categoriali del linguaggio si sovrappongono alle strutture del substrato.

Quest'ultimo punto è la struttura formale della competenza esperta come sincronia substrato-linguaggio: l'esperto non è soltanto colui che ha un substrato ricco e un linguaggio ricco separatamente, ma colui nel quale le strutture categoriali del linguaggio si sono allineate alle strutture del substrato attraverso la co-evoluzione. La perdita della codifica Φ^* è minimizzata per il dominio dell'expertise non perché il codice sia diventato più potente in assoluto, ma perché la geometria di $\mathbb{R}^{n*}$ si è adattata alla struttura di L nel dominio rilevante.

La conseguenza formale è che il punto fisso $(m_{\text{sub}}, m_{\text{lang}})$ dipende dal percorso di co-evoluzione: dalla sequenza di esperienze $e(t)$ e di input comunicativi $c(t)$ attraverso cui l'agente ha attraversato lo spazio degli stati. Due sommelier con la stessa esperienza percettiva ma formati in tradizioni linguistiche diverse — il vocabolario della critica enologica francese versus quello anglosassone, per esempio — convergeranno a punti fissi con geometrie di $\mathbb{R}^{n*}$ diversamente strutturate, riflettendo la diversa partizione categoriale dei rispettivi vocabolari.

VI.9 Il linguaggio interno come operatore sul substrato: il sé che si auto-rimodella

Le analisi precedenti hanno considerato il linguaggio esterno — la comunicazione con altri agenti — come input al sistema dinamico. Ma la componente più sottile e pervasiva della modifica del substrato attraverso il linguaggio è quella operata dal linguaggio interno: il dialogo che l'agente intrattiene con se stesso, il pensiero verbale, la narrazione interiore.

Il linguaggio interno — in senso ampio, tutto il pensiero verbalizzato che non viene comunicato all'esterno — produce gli stessi effetti sulla geometria di $\mathbb{R}^n$ del linguaggio esterno, con alcune differenze strutturalmente importanti. A differenza del linguaggio esterno, il linguaggio interno è generato dallo stesso sistema che lo riceve: m_{lang} al tempo t produce proposizioni che vengono immediatamente decodificate attraverso Φ^{-1} e aggiornano $W_{\{t\}}$. Il circuito è chiuso su se stesso, e la dinamica risultante è un sistema a retroazione che può amplificare o attenuare le strutture esistenti del substrato.

Formalmente, si definisce la componente di auto-retroazione del sistema dinamico come:

$$dm_{\text{sub}}/dt|_{\text{interno}} = f_{\text{lang}}(m_{\text{sub}}, m_{\text{lang}}) = \int_{\mathcal{L}} p(p | m_{\text{lang}}) \cdot \nabla_W \mathcal{L}(p, W) dp$$

dove $p(p | m_{\text{lang}})$ è la distribuzione di probabilità sulle proposizioni generate dalla competenza linguistica corrente m_{lang} . Questa equazione dice che lo stato del substrato è continuamente aggiornato in direzione delle strutture corrispondenti alle proposizioni più probabili sotto la distribuzione linguistica corrente. Il sé che si racconta — che si descrive, si critica, si incita, si consola attraverso il linguaggio interno — sta continuamente rimodellando il proprio substrato.

Questo meccanismo è il fondamento formale di fenomeni psicologici ben documentati: gli effetti dell'auto-talk sulla performance sportiva, gli effetti della narrazione interiore sulla regolazione emotiva, gli effetti del pensiero ruminativo sull'amplificazione degli stati ansiosi. In tutti questi casi, il linguaggio interno agisce come operatore su $\mathbb{R}^n$: la ruminazione ansiosa non è semplicemente il riflesso di uno stato ansioso — è un operatore che mantiene e amplifica le strutture del substrato corrispondenti all'ansia attraverso la retroazione continua tra m_{lang} e m_{sub} .

Il punto fisso del sistema sotto auto-retroazione linguistica pura — senza input di esperienza $e(t)$ né di comunicazione esterna $c(t)$ — è determinato esclusivamente dalla struttura di m_{lang} : il substrato converge alle strutture che la competenza linguistica corrente descrive e ri-descrive. Questo è il meccanismo dell'autorealizzazione delle narrazioni identitarie: il soggetto che si narra come incapace, ansioso, indegno orienta il proprio substrato verso strutture che rendono quelle descrizioni sempre più accurate — non per un processo mistico, ma per il meccanismo preciso dell'aggiornamento sinaptico indotto dal linguaggio interno.

Questo capitolo si connette direttamente con la monografia sul malinteso intrapersonale. Il "sé che non si capisce" — il soggetto la cui narrazione interiore diverge sistematicamente dalla struttura del proprio substrato — è un sistema lontano dall'equilibrio del punto fisso: m_{sub} e m_{lang} sono disallineati, la co-evoluzione procede lungo una traiettoria instabile, e la perdita informazionale di $\Phi(m_{\text{sub}})$ è alta perché la geometria del substrato non si è orientata verso la struttura del linguaggio disponibile. Il lavoro terapeutico — in molte delle sue forme — può essere formalizzato come l'intervento sul termine $c(t)$ dell'equazione di m_{lang} : la fornitura di un vocabolario più ricco e più accurato per le strutture del substrato, che avvia una co-evoluzione verso un nuovo punto fisso con perdita di Φ più bassa.

VI.10 Memoria come narrazione, riconsolidamento come riscrittura: il ciclo completo

L'analisi del sistema dinamico accoppiato trova la propria espressione più netta nel fenomeno del riconsolidamento della memoria — il processo attraverso cui i ricordi, ogni volta che vengono richiamati, vengono modificati dalla struttura cognitiva corrente prima di essere re-immagazzinati. Il riconsolidamento è il punto in cui il ciclo substrato→linguaggio→substrato si chiude con maggiore evidenza empirica e con le conseguenze più rilevanti per la teoria dell'identità personale.

Si formalizza il processo di riconsolidamento nel quadro dinamico accoppiato. Sia $v_m(t_0)$ il vettore del substrato che codifica un ricordo al tempo della sua formazione t_0 . Al tempo $t_1 > t_0$, quando il ricordo viene richiamato, il sistema produce non $v_m(t_0)$ ma una ricostruzione $v'_m(t_1)$ attraverso un processo di decodifica-ricodifica:

$$v'_m(t_1) = \Phi^{-1}_{\{t_1\}}(\Phi_{\{t_1\}}(v_m(t_0)))$$

dove $\Phi_{\{t_1\}}$ e $\Phi^{-1}_{\{t_1\}}$ sono le funzioni di codifica e decodifica allo stato del sistema al tempo t_1 . Poiché la geometria di $\mathbb{R}^n_{\{t_1\}}$ differisce da quella di $\mathbb{R}^n_{\{t_0\}}$ a causa della co-evoluzione del sistema negli intervalli $[t_0, t_1]$, la ricostruzione $v'_m(t_1)$ è strutturalmente diversa dal vettore originale $v_m(t_0)$. La distanza:

$$d_ricons = 1 - \text{sim}(v_m(t_0), v'_m(t_1))$$

misura la distorsione del riconsolidamento — quanto il ricordo si è allontanato dalla struttura originale attraverso la ricostruzione. Questa distanza è tanto maggiore quanto più la geometria di $\mathbb{R}^n$ è cambiata tra t_0 e t_1 .

Il punto cruciale per il ciclo completo substrato-linguaggio-substrato è che la narrativizzazione del ricordo — il racconto che l'agente produce del proprio passato — non è soltanto il riflesso del riconsolidamento: è un operatore che lo accelera e lo dirige. Quando un agente narra un ricordo in una certa forma narrativa — selezionando gli aspetti enfaticizzati, inquadrandoli in una struttura causale, collocandoli in una trama identitaria — attiva le funzioni Φ e Φ^{-1} sulle proposizioni della narrazione, e queste proposizioni aggiornano la geometria di $\mathbb{R}^n$ in direzione della struttura narrativa adottata. Il ricordo che viene raccontato in un certo modo tende a diventare, nel substrato, un ricordo con la struttura del racconto. La memoria come finzione non è una metafora: è la struttura precisa del processo di riconsolidamento attraverso la narrativizzazione, formalizzata come trasformazione iterata di $\mathbb{R}^n$ attraverso le proposizioni del racconto.

Il ciclo completo del sistema dinamico accoppiato è pertanto:

$$\text{Esperienza} \rightarrow v \in \mathbb{R}^n \rightarrow \Phi \rightarrow \text{narrazione in } L \rightarrow \Phi^{-1} \rightarrow v' \in \mathbb{R}^n \rightarrow \text{modifica di } \mathbb{R}^n \rightarrow \text{nuova esperienza percepita attraverso } \mathbb{R}^n \text{ modificato} \rightarrow \dots$$

Ogni giro del ciclo può avvicinarsi al punto fisso (convergenza co-evolutiva) o allontanarsene (divergenza amplificata dalla retroazione). L'equilibrio cognitivo — la

struttura stabile di un agente che ha integrato esperienza e narrazione — non è un dato: è il risultato di un processo dinamico che può essere descritto, analizzato e, in linea di principio, guidato. Il confine tra psicologia cognitiva, pedagogia, psicoterapia e teoria della mente si rivela, nel quadro del sistema dinamico accoppiato, come differenza di prospettiva sullo stesso processo formale.

VI.11 Le condizioni al contorno: quando il linguaggio modifica il substrato di più

L'intensità della modifica del substrato indotta dal linguaggio — la norma del termine f_{lang} nell'equazione dinamica — non è costante: dipende da condizioni al contorno che il quadro formale permette di identificare con precisione. Identificare queste condizioni è rilevante tanto per la psicologia dell'apprendimento quanto per la comprensione dei meccanismi attraverso cui interventi linguistici producono effetti cognitivi stabili.

La prima condizione è la *granularità lessicale del dominio*: quanto più fine è il vocabolario disponibile per descrivere un dominio, tanto più forte è l'effetto della verbalizzazione sulla geometria del substrato in quel dominio. La norma di f_{lang} è proporzionale a $|V_D|$, il numero di termini disponibili nel vocabolario D . Un chirurgo che apprende il vocabolario della neuroanatomia — centinaia di termini per strutture distinte — subisce una trasformazione della propria geometria percettiva neuroanatomica di intensità molto superiore a quella di un profano che apprende i dieci termini del vocabolario neuroanatomico ordinario.

La seconda condizione è la *coerenza semantica del vocabolario acquisito*: quanto più sistematica è la struttura semantica del vocabolario — quanto più le relazioni tra i termini riflettono relazioni nel dominio referenziale — tanto più la trasformazione di $\mathbb{R}^n$ indotta dall'acquisizione è strutturalmente coerente e permanente. Un vocabolario incoerente, in cui i termini si sovrappongono o contraddicono, induce trasformazioni instabili di $\mathbb{R}^n$ che tendono a dissiparsi.

La terza condizione è la *fase di apprendimento*: la plasticità del substrato — la velocità con cui $\mathbb{R}^n$ risponde agli aggiornamenti indotti dal linguaggio — è massima nei periodi critici dello sviluppo, decresce nell'età adulta, e tende a un asintoto basso nell'expertise consolidata. Questo corrisponde formalmente alla variazione temporale del tasso di apprendimento η nella regola di aggiornamento: $\eta(t)$ è alta nella fanciullezza e decresce monotonicamente con il consolidamento dell'expertise. Ne consegue che i periodi di maggiore plasticità linguistica — l'acquisizione della lingua madre, i periodi di riqualificazione professionale, i momenti di apertura cognitiva indotta da crisi o rivelazioni — sono anche i periodi in cui il linguaggio ha il massimo potere trasformativo sul substrato.

La quarta condizione è l'*attenzione allocata alla verbalizzazione*: la norma di f_{lang} è proporzionale all'intensità dell'attenzione con cui il soggetto produce e comprende le

proposizioni linguistiche nel dominio. Il linguaggio distratto — la lettura automatica, il parlare senza ascoltarsi — produce aggiornamenti di $\mathbb{R}^n$ deboli e spesso incoerenti. Il linguaggio attento — la scrittura riflessiva, la lettura analitica, il dialogo profondo — produce aggiornamenti forti e strutturalmente coerenti. Questo è il meccanismo formale dell'effetto ben documentato della scrittura espressiva sulla regolazione emotiva e sull'elaborazione del lutto: la narrativizzazione attenta degli stati emotivi ristrutturata il substrato emotivo più efficacemente della sola esperienza emotiva non verbalizzata.

VI.12 Implicazioni per la teoria della cognizione e per i sistemi artificiali

Il sistema dinamico accoppiato substrato-linguaggio produce implicazioni che attraversano i confini tra filosofia della mente, psicologia cognitiva, neuroscienze computazionali e intelligenza artificiale.

Per la filosofia della mente, il quadro risolve la falsa alternativa tra nativismo e costruttivismo nella struttura del substrato cognitivo. Il substrato non è né completamente determinato dall'esperienza sensorimotoria (costruttivismo radicale) né completamente indipendente dall'ambiente culturale e linguistico (nativismo radicale): è un sistema dinamico la cui struttura è co-determinata dall'esperienza del mondo fisico e dall'immersione in un ambiente linguistico-culturale, con il peso relativo delle due sorgenti che varia per dominio, per età, e per tipo di conoscenza.

Per la psicologia cognitiva, il quadro offre una formalizzazione unificata di fenomeni apparentemente eterogenei: il verbal overshadowing, il moral deskilling, l'expertise percettiva, gli effetti dell'auto-talk sulla performance, il riconsolidamento della memoria, l'effetto della psicoterapia narrativa. Tutti questi fenomeni sono istanze del medesimo meccanismo: la modifica del substrato attraverso la retroazione linguistica nel sistema dinamico accoppiato.

Per le neuroscienze computazionali, il quadro identifica il meccanismo molecolare e sinaptico attraverso cui il linguaggio modifica il substrato: la regola di aggiornamento $W_{t+\delta t} = W_t + \eta \cdot \partial \mathcal{L} / \partial W_t$ è una forma generalizzata della plasticità sinaptica Hebbiana. Le previsioni quantitative del modello — sulla velocità di trasformazione di $\mathbb{R}^n$ in funzione della granularità lessicale, sul coefficiente κ dell'effetto Whorf, sulla struttura dei punti fissi dell'expertise — sono in linea di principio verificabili attraverso misure di neuroimaging funzionale e analisi della geometria degli spazi di attivazione.

Per i sistemi artificiali, le implicazioni sono radicali. Il Capitolo IV ha argomentato che i modelli di linguaggio sono sistemi senza substrato, il cui spazio di embedding riflette la struttura distribuzionale del linguaggio anziché la struttura dell'esperienza sensorimotoria. Il quadro dinamico aggiunge una conseguenza ulteriore: anche se un sistema artificiale disponesse di un substrato — come un sistema robotico con sensori e attuatori — il suo sviluppo cognitivo richiederebbe non solo l'esposizione

all'esperienza sensorimotoria ma la co-evoluzione del substrato con un sistema linguistico, attraverso un ciclo di verbalizzazione-rimodellamento che trasformi progressivamente la geometria del substrato in direzione delle strutture concettualmente rilevanti. Un sistema robotico che apprende solo per reinforcement sensorimotorio, senza ciclo linguistico, svilupperà un substrato ricco ma linguisticamente disallineato — una forma estrema del caso del ciclista prima della verbalizzazione. Un sistema che apprende solo attraverso il linguaggio, senza substrato sensorimotorio, svilupperà un sistema linguistico ricco ma subproposizionalmente vuoto — il caso dei modelli di linguaggio attuali. La cognizione umana matura è il risultato della co-evoluzione di entrambi, attraverso un ciclo la cui formalizzazione è il contributo centrale di questo capitolo.

Appendici tecniche

Appendice A — Derivazioni formali del sistema dinamico accoppiato

A.1 Spazio degli stati e metriche

Si formalizza con precisione lo spazio degli stati del sistema dinamico accoppiato. Lo stato del substrato m_{sub} appartiene allo spazio M_{sub} , che si identifica con lo spazio delle matrici di peso $W \in \mathbb{R}^{N \times N}$ modulo equivalenza sinaptica, dove N è il numero di neuroni nel sistema. In pratica, poiché la dinamica cognitiva avviene su varietà a dimensionalità intrinseca molto più bassa di N^2 , si lavora con la proiezione di M_{sub} sulla varietà cognitiva rilevante $M \subset \mathbb{R}^n$ introdotta nel Capitolo I.

La metrica su M_{sub} è la metrica di Frobenius tra matrici di peso, ridotta alla varietà M :

$$d_{\text{sub}}(m_{\text{sub},1}, m_{\text{sub},2}) = \|W_1 - W_2\|_F / \|W_0\|_F$$

normalizzata allo stato basale W_0 . Questa metrica è invariante per riscaldamento globale dei pesi, il che è appropriato poiché sono le strutture relative dei pesi a determinare la geometria della rappresentazione, non i valori assoluti.

Lo stato linguistico m_{lang} appartiene allo spazio M_{lang} , che si identifica con lo spazio delle distribuzioni di probabilità su $L \times \mathcal{C}(L)$, dove $\mathcal{C}(L)$ è lo spazio dei contesti linguistici. In pratica, M_{lang} è caratterizzato da: un insieme di parametri lessicali (il vocabolario attivo e le soglie di applicabilità di ogni termine), un insieme di parametri sintattici (le distribuzioni di struttura sintattica), e un insieme di parametri pragmatici

(le regole di adeguatezza contestuale). La metrica su M_{lang} è la distanza di Kullback-Leibler tra distribuzioni:

$$d_{\text{lang}}(m_{\text{lang},1}, m_{\text{lang},2}) = D_{\text{KL}}(p_1 \parallel p_2) = \sum_{\{p \in L\}} p_1(p) \cdot \log(p_1(p) / p_2(p))$$

che misura la divergenza tra le distribuzioni di produzione linguistica dei due stati.

A.2 Le funzioni f e g : struttura analitica

La funzione $f : M_{\text{sub}} \times M_{\text{lang}} \times \mathcal{E} \rightarrow T_{\{m_{\text{sub}}\}}(M_{\text{sub}})$ che governa l'evoluzione del substrato si decompone analiticamente come:

$$f(m_{\text{sub}}, m_{\text{lang}}, e) = f_{\text{exp}}(m_{\text{sub}}, e) + f_{\text{lang}}(m_{\text{sub}}, m_{\text{lang}}) + f_{\text{int}}(m_{\text{sub}}, m_{\text{lang}}, e)$$

Il termine f_{exp} descrive la plasticità sinaptica indotta dall'esperienza sensorimotoria pura — il meccanismo Hebbiano standard. Per un vettore di esperienza $e(t)$ che produce l'attivazione $v_e = v_e(t)$ nel substrato, si ha:

$$f_{\text{exp}}(m_{\text{sub}}, e) = \eta_{\text{exp}} \cdot (v_e \otimes v_e - W \cdot v_e \otimes v_e)$$

dove $v_e \otimes v_e$ è il prodotto esterno del vettore di attivazione con se stesso (regola di Hebb), e il secondo termine è il termine di decadimento (plasticità omeostatica). Questa equazione è la forma differenziale della regola di Hebb con decadimento sinaptico.

Il termine f_{lang} descrive la plasticità sinaptica indotta dall'elaborazione linguistica. Per una proposizione p prodotta o compresa al tempo t con vettore di decodifica $v_p = \Phi^{-1}(p)$:

$$f_{\text{lang}}(m_{\text{sub}}, m_{\text{lang}}) = \eta_{\text{lang}} \cdot \mathbb{E}_{\{p \sim p(\cdot \mid m_{\text{lang}})\}} [v_p \otimes v_p - W \cdot v_p \otimes v_p]$$

dove la media è presa sulla distribuzione di proposizioni prodotte dalla competenza linguistica corrente. Questo è il termine chiave assente nel modello unidirezionale: la plasticità sinaptica indotta non dall'esperienza diretta ma dalla struttura attesa della produzione linguistica.

Il termine f_{int} è il termine di interazione: descrive l'effetto di sinergia tra esperienza e linguaggio quando entrambi co-attivano le stesse strutture del substrato. L'apprendimento è più rapido quando l'esperienza e la verbalizzazione dell'esperienza avvengono in stretta prossimità temporale:

$$f_{\text{int}}(m_{\text{sub}}, m_{\text{lang}}, e) = \eta_{\text{int}} \cdot \mathbb{E}[(v_e \otimes v_p + v_p \otimes v_e) \cdot 1_{\{d(v_e, v_p) < \delta\}}]$$

dove $1_{\{d(v_e, v_p) < \delta\}}$ è la funzione indicatrice dell'evento che il vettore di esperienza e il vettore di decodifica della proposizione siano sufficientemente vicini nello spazio di attivazione.

La funzione $g : M_{\text{sub}} \times M_{\text{lang}} \times \mathcal{C} \rightarrow T_{\{m_{\text{lang}}\}}(M_{\text{lang}})$ che governa l'evoluzione della competenza linguistica è:

$$g(m_{\text{sub}}, m_{\text{lang}}, c) = g_{\text{sub}}(m_{\text{sub}}, m_{\text{lang}}) + g_{\text{comm}}(m_{\text{lang}}, c)$$

Il termine g_{sub} descrive l'adattamento della competenza linguistica alla struttura del substrato. Formalmente, è il gradiente della perdita di ricostruzione del substrato attraverso la codifica linguistica:

$$g_{\text{sub}}(m_{\text{sub}}, m_{\text{lang}}) = -\eta_{\text{vocab}} \cdot \nabla_{\{m_{\text{lang}}\}} \mathbb{E}_{\{v \sim q(m_{\text{sub}})\}} [d_{\text{sub}}(v, \Phi^{-1}(\Phi(v)))]$$

dove $q(m_{\text{sub}})$ è la distribuzione di probabilità sugli stati del substrato. Questa equazione dice che la competenza linguistica evolve verso la configurazione che minimizza la distorsione di codifica del substrato: il vocabolario si adatta alla struttura del substrato nel senso di ridurre la perdita informazionale.

Il termine g_{comm} descrive l'adattamento della competenza linguistica all'input comunicativo — l'acquisizione di vocabolario e strutture sintattiche per esposizione al linguaggio altrui:

$$g_{\text{comm}}(m_{\text{lang}}, c) = \eta_{\text{comm}} \cdot \nabla_{\{m_{\text{lang}}\}} \mathbb{E}_{\{p \sim c(t)\}} [\log p(p \mid m_{\text{lang}})]$$

il gradiente della log-verosimiglianza delle proposizioni ricevute rispetto ai parametri della competenza linguistica: la regola di massima verosimiglianza per l'aggiornamento della distribuzione linguistica.

A.3 Analisi della stabilità lineare

Per analizzare la stabilità dei punti fissi, si assume che f e g siano differenziabili in un intorno di $(m_{\text{sub}}, m_{\text{lang}})$ con esperienza e comunicazione medie $(\bar{e}, \bar{c})$. La Jacobiana del sistema linearizzato ha la struttura a blocchi descritta nel testo principale. Si calcolano ora i blocchi esplicitamente.

Il blocco $J_{11} = \partial f / \partial m_{\text{sub}}$ è dominato dal termine di decadimento sinaptico di f_{exp} . In notazione semplificata scalare, si ha:

$$J_{11} \approx -\eta_{\text{exp}} \cdot I + \partial(\eta_{\text{exp}} \cdot v_e \otimes v_e) / \partial W \mid_{\{W^*\}}$$

Il primo termine è la matrice di decadimento sinaptico — negativa definita, garante della stabilità locale nella direzione del substrato in assenza di input. Il secondo termine è la perturbazione indotta dalla struttura dell'esperienza media $\bar{e}$.

Il blocco $J_{12} = \partial f / \partial m_{\text{lang}}$ è il blocco che formalizza l'effetto Whorf. Si ha:

$$J_{12} = \eta_{\text{lang}} \cdot \partial / \partial m_{\text{lang}} \mathbb{E}_{\{p \sim p(\cdot | m_{\text{lang}})\}} [v_p \otimes v_p] |_{\{m^*_{\text{lang}}\}}$$

Questo blocco è non nullo se e soltanto se la distribuzione di produzione linguistica $p(\cdot | m_{\text{lang}})$ cambia con m_{lang} — il che è certamente vero per qualsiasi competenza linguistica non degenera. La norma di J_{12} è proporzionale alla varianza della distribuzione sulle decodifiche $\Phi^{-1}(p)$ pesate per la sensibilità di $p(\cdot | m_{\text{lang}})$ ai parametri m_{lang} . In pratica, $|J_{12}|$ è grande nei domini in cui il vocabolario è ricco e variabile, e piccola nei domini in cui il vocabolario è povero o fisso.

Il blocco $J_{21} = \partial g / \partial m_{\text{sub}}$ è:

$$J_{21} = -\eta_{\text{vocab}} \cdot \partial / \partial m_{\text{sub}} \nabla_{\{m_{\text{lang}}\}} \mathbb{E}_{\{v \sim q(m_{\text{sub}})\}} [d_{\text{sub}}(v, \Phi^{-1}(\Phi(v)))]$$

il cui valore dipende dalla sensibilità della distorsione di codifica alla struttura del substrato. Questo blocco è generalmente non nullo: cambiamenti nella geometria del substrato modificano l'errore di ricostruzione e quindi la direzione di adattamento della competenza linguistica.

Il blocco $J_{22} = \partial g / \partial m_{\text{lang}}$ è dominato dal termine di g_{comm} e riflette la stabilità della competenza linguistica in risposta a perturbazioni dei propri parametri: la tendenza delle abitudini linguistiche a mantenersi stabili in assenza di input comunicativi perturbanti.

La condizione necessaria e sufficiente per la stabilità del punto fisso è che tutti gli autovalori λ_i della matrice J abbiano parte reale negativa: $\text{Re}(\lambda_i) < 0$ per ogni i . Poiché J è una matrice a blocchi con struttura:

$$J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}$$

una condizione sufficiente per la stabilità è data dal criterio di Schur: il punto fisso è stabile se J_{11} e $J_{22} - J_{21} \cdot J_{11}^{-1} \cdot J_{12}$ sono entrambe stabili (tutti gli autovalori con parte reale negativa). Intuitivamente: il punto fisso è stabile se sia il substrato sia la competenza linguistica sono individualmente stabili, e se l'accoppiamento tra i due sistemi non è troppo forte da destabilizzare il sistema complessivo.

A.4 Condizioni di biforcazione

Quando la norma dell'accoppiamento $|J_{12}| \cdot |J_{21}|$ supera una soglia critica determinata da J_{11} e J_{22} , il punto fisso può diventare instabile attraverso una biforcazione. Le due tipologie di biforcazione rilevanti per il sistema substrato-linguaggio sono:

La *biforcazione nodo-sella* avviene quando un autovalore reale di J attraversa lo zero. In termini cognitivi, corrisponde alla transizione da un regime con un unico attrattore cognitivo stabile a un regime con due attrattori stabili separati da una sella instabile — la struttura della "svolta cognitiva" in cui un agente acquisisce una nuova prospettiva concettuale incompatibile con quella precedente. Il cambio di paradigma scientifico, la conversione religiosa, la ristrutturazione terapeutica dell'identità narrativa sono possibili istanze di questa biforcazione.

La *biforcazione di Hopf* avviene quando una coppia di autovalori complessi coniugati attraversa l'asse immaginario. In termini cognitivi, corrisponde alla comparsa di oscillazioni stabili nel sistema substrato-linguaggio: cicli in cui il substrato porta verso strutture che il linguaggio non può esprimere, inducendo adattamenti linguistici che a loro volta rimodellano il substrato in nuove direzioni, in un ciclo che non converge a un punto fisso ma a un attrattore periodico. Questo potrebbe corrispondere alle oscillazioni tra pensiero visivo e pensiero verbale che alcuni matematici e artisti descrivono come caratteristiche del proprio processo creativo.

Appendice B — Fondamenti di teoria della misura per lo spazio del substrato

B.1 La misura su $\mathbb{R}^n$

Lo spazio di attivazione $\mathbb{R}^n$ è uno spazio euclideo finito-dimensionale. Per le applicazioni cognitive, si lavora con la misura di probabilità q su $\mathbb{R}^n$ che rappresenta la distribuzione degli stati cognitivi di un agente in condizioni normali di attività. Questa misura non è la misura di Lebesgue uniforme su $\mathbb{R}^n$: gli stati cognitivi non sono distribuiti uniformemente nello spazio di attivazione, ma si concentrano su sottoinsiemi di misura di Lebesgue zero — le varietà $M \subset \mathbb{R}^n$ introdotte nel Capitolo I.

Si assume che q sia una misura assoluta continua rispetto alla misura di Hausdorff d -dimensionale μ_H su M , dove $d = d_{\text{int}}$ è la dimensionalità intrinseca della varietà cognitiva. La densità di probabilità è:

$$dq / d\mu_H = p_{\text{cog}} : M \rightarrow \mathbb{R}_{>0}$$

una funzione positiva sulla varietà. La struttura di p_{cog} riflette la struttura dell'apprendimento passato: le regioni di M frequentemente visitate dall'agente nel corso della propria storia cognitiva hanno densità p_{cog} alta; le regioni rare hanno densità bassa.

B.2 La varietà cognitiva e le sue proprietà di regolarità

La varietà M su cui gli stati cognitivi sono distribuiti ha proprietà di regolarità che dipendono dalla storia dell'agente. Per un agente novizio in un dominio, la varietà è relativamente piatta e poco strutturata — la distribuzione p_{cog} è quasi uniforme sulla regione del dominio. Per un agente esperto, la varietà ha curvatura concentrata nelle regioni corrispondenti alle strutture categoriali dell'expertise: alta curvatura nelle vicinanze dei prototipi categoriali, bassa curvatura nelle regioni inter-categoriali.

La curvatura di M in un punto $v \in M$ è misurata dal tensore di curvatura riemanniana $R_{\{ijkl\}}(v)$, le cui componenti principali misurano la rapidità con cui la varietà si piega nelle diverse direzioni. La curvatura è massima nei prototipi categoriali e riflette la struttura delle distinzioni finì che il substrato realizza in quelle regioni. La norma del tensore di curvatura:

$$\kappa(v) = (\sum_{\{i,j,k,l\}} R^2_{\{ijkl\}}(v))^{1/2}$$

è una misura scalare della curvatura locale. Si può dimostrare che $\kappa(v)$ è correlata positivamente con la densità $p_{\text{cog}}(v)$: le regioni ad alta densità (stati cognitivi frequenti) tendono ad avere alta curvatura, riflettendo la struttura fine delle distinzioni categoriali apprese.

B.3 La funzione Φ come mappa misurabile

La funzione di codifica linguistica $\Phi : \mathbb{R}^n \rightarrow L$ deve essere trattata come una funzione misurabile rispetto alla σ -algebra Boreliana su $\mathbb{R}^n$ e alla σ -algebra discreta su L . Le proprietà di misurabilità di Φ sono garantite dall'assunzione che le regioni $R_p = \Phi^{-1}(p) \subset \mathbb{R}^n$ siano insiemi misurabili — una condizione che esclude le partizioni patologiche di Vitali e che è soddisfatta da qualsiasi funzione di codifica linguistica realizzata da un processo neurale finito.

La perdita informazionale della codifica Φ sotto la misura q è:

$$I_{\text{lost}}(\Phi, q) = H_q(V) - H_q(V \mid \Phi(V)) = I_q(V; \Phi(V))$$

dove $H_q(V)$ è l'entropia differenziale dello stato cognitivo sotto la misura q , e $I_q(V; \Phi(V))$ è l'informazione mutua tra lo stato del substrato e il suo codice linguistico. Questa quantità dipende sia dalla scelta di Φ sia dalla distribuzione q : la stessa funzione di codifica può avere perdita alta per una distribuzione q (agente con substrato ricco in regioni di alta distorsione) e perdita bassa per un'altra distribuzione q' (agente con substrato concentrato in regioni di bassa distorsione).

La funzione di tasso-distorsione $R_{\text{sub}}(D)$ del Capitolo III si esprime in termini di teoria della misura come:

$$R_{\text{sub}}(D) = \min_{\{\Phi : \mathbb{R}^n \rightarrow L, \mathbb{E}_q[d_{\text{sub}}(V, \Phi^{-1}(\Phi(V)))] \leq D\}} I_q(V; \Phi(V))$$

il minimo dell'informazione mutua tra substrato e codice linguistico soggetto al vincolo che la distorsione media sia al più D . Per valori piccoli di D , questa funzione cresce rapidamente: ridurre la distorsione della codifica richiede aumenti proporzionalmente maggiori dell'informazione trasmessa.

Appendice C — Fondamenti di teoria dell'informazione per la codifica linguistica

C.1 La teoria di Shannon applicata al canale substrato-linguaggio

La relazione tra substrato subpropositivo e linguaggio è formalmente un problema di codifica sorgente con distorsione. La sorgente è la variabile aleatoria V con valori in $\mathbb{R}^n$ e distribuzione q ; il codice è lo spazio discreto L dei simboli linguistici; la distorsione è la distanza angolare $d_{\text{sub}} = 1 - \text{sim}(v, \Phi^{-1}(\Phi(v)))$.

Il teorema di tasso-distorsione di Shannon stabilisce che, per distorsione D , la quantità minima di informazione che deve essere trasmessa dal codice è $R(D)$ — la funzione di tasso-distorsione. Nel caso della codifica linguistica, questa funzione ha la forma:

$$R_{\text{sub}}(D) = H(V) - \max_{\{p(p|v) : \mathbb{E}[d_{\text{sub}}] \leq D\}} I(V; P)$$

dove il massimo è preso sull'insieme delle distribuzioni condizionali $p(p|v)$ di proposizioni dato lo stato del substrato che soddisfano il vincolo di distorsione. Per sorgenti gaussiane con metrica quadratica, la funzione $R(D)$ ha la forma analitica nota:

$$R(D) = (1/2) \cdot \log_2(\sigma^2_V / D) \quad \text{per } D \leq \sigma^2_V$$

dove σ^2_V è la varianza della sorgente. Per la sorgente cognitiva con struttura non gaussiana, la funzione $R_{\text{sub}}(D)$ deve essere calcolata numericamente, ma la struttura qualitativa è analoga: cresce logicamente per distorsione decrescente, con un punto di saturazione per $D \rightarrow 0$ dove $R_{\text{sub}}(D) \rightarrow H(V)$.

C.2 L'informazione mutua come misura della co-dipendenza substrato-linguaggio

Nel sistema dinamico accoppiato, l'informazione mutua $I(M_{\text{sub}} ; M_{\text{lang}})$ tra le variabili di stato del substrato e della competenza linguistica è una misura naturale del grado di co-dipendenza delle due variabili. Questa quantità evolve nel tempo secondo l'equazione:

$$dI(M_{\text{sub}}(t) ; M_{\text{lang}}(t))/dt = \partial I / \partial t|_{\text{f}} + \partial I / \partial t|_{\text{g}}$$

dove il primo termine è il contributo all'informazione mutua indotto dall'evoluzione del substrato e il secondo è il contributo indotto dall'evoluzione della competenza linguistica. In uno stato stazionario (punto fisso), questa derivata è zero: le due variabili di stato hanno raggiunto un livello di dipendenza mutua stabile.

Il valore dell'informazione mutua allo stato stazionario $I^*(M_{\text{sub}} ; M_{\text{lang}})$ misura il grado di allineamento raggiunto dalla co-evoluzione: quanto la struttura del substrato è predittiva della struttura linguistica e viceversa. Per l'esperto maturo, I^* è alta nel dominio dell'expertise: la geometria di $\mathbb{R}^n$ nel dominio riflette fedelmente la struttura del vocabolario tecnico, e viceversa. Per il novizio, I^* è bassa nello stesso dominio: substrato e linguaggio sono quasi indipendenti.

C.3 La capacità del canale linguistico e il teorema di incompressibilità

Un risultato fondamentale che il quadro formale permette di derivare è il teorema di incompressibilità della conoscenza procedurale esperta, presentato informalmente nel Capitolo V. Si enuncia qui nella forma precisa.

Si consideri un dominio cognitivo D con dimensionalità intrinseca d_{int} e un vocabolario linguistico L_D con $|L_D|$ termini. La capacità del canale linguistico per il dominio D è:

$$C_D = \log_2 |L_D| \quad \text{bit}$$

La struttura dell'expertise nel dominio D ha entropia:

$$H_{\text{expert}}(D) = d_{\text{int}} \cdot \log_2(1/\epsilon) \quad \text{bit}$$

dove ϵ è la precisione con cui l'expertise discrimina tra stati diversi del substrato nel dominio D . Per l'expertise motoria fine — il pianista, il chirurgo, il ceramista — ϵ è dell'ordine di 10^{-3} rad di angolo nello spazio del movimento, e d_{int} è dell'ordine di 10^2 - 10^3 per il sistema corpo umano completo.

Il teorema di incompressibilità afferma: per qualsiasi codifica linguistica Φ_D della conoscenza procedurale esperta nel dominio D , la distorsione è:

$$D_{\text{min}}(D) \geq 1 - 2^{\{-(H_{\text{expert}}(D) - C_D)\} / \sigma_V}$$

dove σ_V è la deviazione standard dell'attivazione nel dominio D . La distorsione è positiva e crescente con $H_{\text{expert}}(D) - C_D = d_{\text{int}} \cdot \log_2(1/\epsilon) - \log_2 |L_D|$: quanto più la ricchezza del substrato supera la capacità del vocabolario, tanto maggiore è la perdita irriducibile.

Per la conoscenza procedurale del pianista esperto: $d_{\text{int}} \approx 10^3$ (dimensionalità dello spazio dei movimenti pianistici rilevanti), $\log_2(1/\epsilon) \approx 10$ bit (precisione milliradiana), e $|L_D| \approx 10^3$ (termini del vocabolario tecnico pianistico). Si ottiene $H_{\text{expert}} \approx 10^4$

bit e $C_D \approx 10$ bit — un rapporto di compressione superiore a $10^3:1$. Il teorema prevede che qualsiasi manuale di tecnica pianistica codifichi meno dello 0.1% della struttura dell'expertise procedurale del pianista esperto.

Appendice D — Connessioni con la letteratura empirica sul relativismo linguistico

D.1 Il programma empirico di Winawer et al.

Lo studio di Winawer et al. (2007) ha confrontato la discriminabilità cromatica di parlanti russo e inglese per coppie di colori che attraversano la frontiera categoriale blu chiaro/blu scuro (differenziata in russo dai termini "goluboy" e "siniy", non distinta in inglese). I risultati mostrano che i parlanti russo discriminano più rapidamente le coppie inter-categoriali — quelle che attraversano la frontiera nominale — rispetto alle coppie intra-categoriali di uguale distanza fisica nello spazio di Munsell.

Nel quadro formale del sistema dinamico accoppiato, questo risultato corrisponde all'effetto del blocco J_12 sulla geometria di $\mathbb{R}^n_{\{\text{visivo}\}}$: l'acquisizione del termine "goluboy" ha indotto una trasformazione metrica dello spazio cromatico del substrato dei parlanti russo, avvicinando i punti all'interno della categoria e allontanando i punti al confine inter-categoriale. Il coefficiente κ dell'ipotesi di Whorf per il dominio cromatico in russo-inglese è stimabile dai dati di Winawer et al. come:

$$\kappa \approx \Delta \text{SNR} / \Delta G(\text{cromatico, russo, inglese}) \approx 0.15 / 2 \approx 0.075$$

dove ΔSNR è la differenza nel rapporto segnale-rumore della discriminazione cromatica e $\Delta G = 2$ è la differenza di granularità lessicale (2 termini in russo vs 1 in inglese per il blu).

D.2 Lo studio di Boroditsky et al. sulla struttura temporale

Boroditsky et al. (2011) hanno confrontato la struttura cognitiva del tempo in parlanti mandarino e inglese. Il mandarino usa prevalentemente metafore spaziali verticali per il tempo (il tempo futuro è "in basso", il passato è "in alto"), mentre l'inglese usa metafore orizzontali (il futuro è "davanti", il passato è "dietro"). I risultati mostrano che i parlanti mandarino sono più veloci nel rispondere a domande temporali dopo prime verticali rispetto a prime orizzontali, mentre i parlanti inglese mostrano il pattern inverso.

Nel quadro formale, questo corrisponde a una differenza nella geometria della varietà temporale $M_{\{\text{tempo}\}} \subset \mathbb{R}^n_{\{\text{concettuale}\}}$: la struttura delle relazioni di prossimità tra stati cognitivi temporali è orientata lungo l'asse verticale per i parlanti mandarino e lungo l'asse orizzontale per i parlanti inglese, con i due orientamenti che riflettono le strutture metaforiche dominanti nei rispettivi vocabolari temporali. L'effetto è una conseguenza del termine f_{lang} nell'equazione del substrato: il vocabolario metaforico

della temporalità ha orientato la geometria del substrato concettuale nella direzione della struttura spaziale più frequentemente attivata dalla metafora.

D.3 Lo studio di Roberson et al. sulla categorizzazione cromatica trans-culturale

Roberson et al. (2005) hanno confrontato la categorizzazione cromatica di parlanti inglese, Berinmo (lingua della Papua Nuova Guinea) e Himba (lingua della Namibia), mostrando che le frontiere tra categorie di colore variano tra le tre lingue e che la discriminabilità cromatica segue le frontiere categoriali della propria lingua piuttosto che frontiere universali nello spazio di Munsell.

Nel quadro formale, questo studio fornisce la stima più diretta del coefficiente κ dell'ipotesi di Whorf per il dominio cromatico. La correlazione tra granularità lessicale differenziale ΔG e differenza di discriminabilità ΔSNR per i tre sistemi linguistici-culturali è positiva e significativa, con $\kappa \approx 0.08-0.12$ per il dominio cromatico. Questo coefficiente non è zero — confermando che il blocco J_12 della Jacobiana è non nullo — ma è piccolo rispetto all'effetto totale della struttura cromatica, consistentemente con la versione debole del relativismo linguistico.

Bibliografia

Le opere qui elencate sono quelle su cui il quadro formale della monografia si fonda direttamente, per quanto riguarda i risultati empirici citati e le tradizioni teoriche con cui il quadro si confronta.

Fondamenti di rappresentazione distribuita e spazi di attivazione

Hebb, D.O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York. Il testo che introduce il principio di plasticità sinaptica correlazionale ("cells that fire together, wire together") — la base biologica del termine f_{exp} nell'equazione del substrato.

Rumelhart, D.E., McClelland, J.L., e il PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I e II. MIT Press, Cambridge MA. Il testo fondativo della modellizzazione della cognizione come elaborazione distribuita in spazi di attivazione — il framework che il Capitolo I estende formalizzando la geometria di $\mathbb{R}^n$.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26. L'articolo che introduce word2vec e dimostra empiricamente che le analogie concettuali corrispondono a trasformazioni lineari negli spazi di embedding — la formalizzazione dell'esempio re - uomo + donna $\approx$ regina citato nel Capitolo I.

Conoscenza procedurale e incomprimibilità

Polanyi, M. (1966). *The Tacit Dimension*. Doubleday, Garden City NY. L'opera che introduce il concetto di "conoscenza tacita" — la conoscenza che non si può trasmettere interamente attraverso il linguaggio proposizionale. Il Capitolo V fornisce la formalizzazione di questa tesi in termini di teoria dell'informazione.

Anderson, J.R. (1983). *The Architecture of Cognition*. Harvard University Press, Cambridge MA. Il testo che stabilisce la distinzione dichiarativo/procedurale nel framework ACT* — la distinzione che il Capitolo V traduce nella differenza tra conoscenza in L e conoscenza in $\mathbb{R}^n$.

Fitts, P.M., Posner, M.I. (1967). *Human Performance*. Brooks/Cole, Belmont CA. Il testo classico sulle fasi dell'acquisizione delle competenze motorie — la progressione da un controllo cognitivo-verbale esplicito a un controllo automatico-procedurale implicito, che nel quadro formale corrisponde alla progressione dal dominio di L al dominio di $\mathbb{R}^n$.

Relativismo linguistico e effetto Whorf

Whorf, B.L. (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*, ed. John B. Carroll. MIT Press, Cambridge MA. La raccolta degli scritti in cui Whorf articola l'ipotesi del relativismo linguistico nella sua formulazione originale — la tesi che il Capitolo VI formalizza in termini del coefficiente κ del sistema dinamico.

Winawer, J., Witthoft, N., Frank, M.C., Wu, L., Wade, A.R., Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19), 7780-7785. L'articolo che dimostra empiricamente l'effetto delle categorie cromatiche linguistiche sulla discriminabilità percettiva — la principale evidenza empirica del blocco J_12 non nullo citata nell'Appendice D.

Boroditsky, L., Fuhrman, O., McCormick, K. (2011). Do English and Mandarin speakers think about time differently? *Cognition*, 118(1), 123-129. Lo studio che dimostra la dipendenza della struttura cognitiva del tempo dal sistema di metafore spaziali della propria lingua — un'istanza dell'effetto del termine f_{lang} sui domini concettuali astratti.

Roberson, D., Davies, I., Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129(3), 369-398. Lo studio cross-culturale che fornisce le stime del coefficiente κ per il dominio cromatico tra sistemi linguistici con granularità cromatica molto diversa.

Roberson, D., Davidoff, J., Davies, I.R.L., Shapiro, L.R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, 50(4), 378-411. L'estensione dello studio precedente che confronta tre sistemi linguistici (inglese, Berinmo, Himba) e quantifica la relazione tra granularità lessicale e discriminabilità cromatica.

Problema del grounding e sistemi artificiali

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3), 335-346. L'articolo che introduce formalmente il grounding problem — il problema di ancorare i simboli di un sistema artificiale a referenti non simbolici — il cui contenuto il Capitolo IV del testo principale traduce nel quadro della funzione di codifica Φ .

Brooks, R.A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1-3), 139-159. L'articolo che propone il paradigma dell'AI embodied come alternativa all'AI simbolica — l'approccio che nel quadro formale corrisponde all'importanza del ciclo percezione-azione per la costruzione di $\mathbb{R}^n$.

Sistemi dinamici e cognizione

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21(5), 615-628. L'articolo che propone di trattare la cognizione come sistema dinamico — il framework che il Capitolo VI estende all'accoppiamento substrato-linguaggio.

Thelen, E., Smith, L.B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge MA. Il testo fondativo della teoria dei sistemi dinamici applicata allo sviluppo cognitivo — la prospettiva teorica in cui il sistema accoppiato del Capitolo VI si inserisce.

Riconsolidamento della memoria e narrazione

Nader, K., Schafe, G.E., LeDoux, J.E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406, 722-726. L'articolo che scopre il fenomeno del riconsolidamento sinaptico — la modifica dei ricordi durante il recupero — che il Capitolo VI formalizza come trasformazione del vettore del substrato attraverso la narrativizzazione.

Hyman, I.E., Loftus, E.F. (1998). Errors in autobiographical memory. *Clinical Psychology Review*, 18(8), 933-947. La revisione della letteratura sull'imprecisione della memoria autobiografica — il corpus che il Capitolo VI connette formalmente alla struttura del riconsolidamento come riscrittura del substrato.

Intuizione morale e ragionamento post hoc

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist model of moral judgment. *Psychological Review*, 108(4), 814-834. L'articolo che propone il modello intuizionista sociale del giudizio morale — la tesi che il ragionamento morale esplicito sia post hoc rispetto all'intuizione, che il Capitolo V connette al framework della codifica lossy.

Verbal overshadowing e interferenza verbale

Schooler, J.W., Engstler-Schooler, T.Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1), 36-71. L'articolo che introduce il fenomeno del verbal overshadowing — la degradazione della memoria visiva indotta dalla verbalizzazione — che il Capitolo VI formalizza come la componente negativa $S(m_{\text{sub}}, m_{\text{lang}})$ nell'equazione del substrato.

Dimensionalità intrinseca degli spazi di attivazione

Tenenbaum, J.B., de Silva, V., Langford, J.C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2319-2323. L'articolo che introduce l'algoritmo Isomap per la stima della dimensionalità intrinseca delle varietà ad alta dimensione — il metodo di riferimento per le stime di d_{int} citate nel Capitolo I.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210-1224. Lo studio che mappa la geometria dello spazio di attivazione della corteccia cerebrale per migliaia di categorie semantiche, fornendo stime empiriche della dimensionalità intrinseca dei substrati cognitivi umani.

Psicoterapia come intervento sul sistema accoppiato

Pennebaker, J.W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, 8(3), 162-166. L'articolo che documenta gli effetti della scrittura espressiva sulla salute psicologica e fisiologica — la principale evidenza empirica degli effetti trasformativi della narrativizzazione attenta del substrato emotivo discussi nel paragrafo VI.11.

Expertise percettiva

Gauthier, I., Tarr, M.J., Anderson, A.W., Skudlarski, P., Gore, J.C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience*, 2(6), 568-573. Lo studio che dimostra la riorganizzazione dell'attivazione corticale con l'acquisizione di expertise percettiva — la base neurobiologica della trasformazione T_{vocab} del substrato percettivo descritta nel paragrafo VI.7.